CHARACTERIZATION OF A UNIQUE HIGH MOBILITY GROUP (HMG) BOX DOMAIN OF MOUSE MAELSTROM

by
Pavol Genzor

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March 17. 2015

ABSTRACT

Maelstrom (MAEL) is a gonad-specific protein that is associated with the piRNA pathway and is involved in silencing of transposable elements (TEs). In the absence of MAEL, the function of the piRNA pathway is perturbed and retrotransposons become expressed, causing infertility. Here, I utilize sequence alignments, tertiary structural analysis and biochemical approaches to characterize MAEL's N-terminal high-mobility group (HMG)-box domain, which is required for function in *Drosophila*. I have compared mouse and fruit fly MAEL HMG-box with sequence specific sex-determining region Y (SRY) HMG-box, and non-sequence specific high mobility group protein B1 (HMGB1) HMG-box-A.

Sequence and structural comparisons revealed novel arrangements of residues and tertiary structural regions ("propeller," "hook") that distinguish MAEL HMG-box from previously described HMG-boxes. These characteristics are highly conserved within vertebrate domains but diverged in invertebrate domains. Gel-shift assays show that MAEL HMG-box does not bind to B-type helical or cytosine methylation modified double-stranded (ds) DNA, but strongly binds to structured DNA four-way junctions. More importantly, MAEL HMG-box binds to RNA. It binds to dsRNA, hairpins, and 4WJs forming stronger complexes with each substrate consecutively. Binding to junctions depends on the conserved arginine residues within the "hook" and "propeller" regions. MAEL HMG-box also binds to large RNA fragments from sequence regions enriched in the MAEL immunoprecipitates and not to RNA that was not enriched.

These results indicate that MAEL HMG-box is an RNA-binding domain with preference for large, structured substrates. Accordingly, MAEL HMG-box may bestow

the RNA-binding capabilities on MAEL protein, allowing for proper selection of target-RNA molecules and their delivery to the piRNA pathway.

Thesis Advisor, First reader: Alex Bortvin, PhD.

Second Reader: David Zappulla, PhD.

PREFACE

Proteins can be viewed as molecular machines that carry out the vast majority of the work in the cell. Whether activated by post-translational modifications, inter-protein associations, or with other molecules (e.g., RNA), proteins are capable of accomplishing diverse molecular tasks. A single protein functional repertoire originates from its domains – regions within its peptide with particular sequence and/or tertiary structure characteristics. The protein domains function in concert to accomplish most commonly single biochemical function. While proteins with various domains demonstrate incredible combinatorial power of the evolutionary process, they can present challenges for scientists trying to decipher functions. This is the case of Maelstrom protein whose biochemical function eluded description for over two decades.

Maelstrom's annotation reveals only two domains: the amino-terminal high-mobility group (HMG)-box domain, followed by a Maelstrom-specific domain (MSD) that has been predicted to form an RNAse H-like fold. The Maelstrom (MAEL) in mice is exclusively expressed in the animal gonads, being first detectable at embryonic stages, concomitant with meiosis. Its deletion during this period, and thereafter, leads to male and female infertility. This is due to defects in the piRNA pathway resulting in failure to regulate transposable elements (TEs), which in turn causes a variety of meiotic problems. How MAEL function during these processes is unknown.

Others studying MAEL refer to it as "one of the most enigmatic proteins" for the diversity of the functions that have been attributed to it in mouse, fruit fly, and cell-culture systems. In the fruit fly, MAEL has been suggested to play roles in the establishment of early-oocyte polarity, organization of microtubule-organizing center,

miRNA regulation, and heterochromatin regulation. In mice, MAEL has been suggested to be involved in meiotic silencing of unsynapsed chromatin and found highly expressed in various tumors. However, by far the largest body of evidence places MAEL in the piRNA pathway as one of the essential members.

The majority of all efforts put into trying to understand MAEL function were focused on interpretation of the phenotypes caused by either its deletion or mutation *in vivo*. While such approaches can provide us with biological contexts, when a large number of variables are present, they can mask the biochemical function of a protein. MAELs enigmatic nature stands as proof of this. In order to unmask the biochemical function of MAEL, in this thesis, I will focus on the characterization of its amino-terminal HMG-box domain. I will present multiple lines of evidence supporting the hypothesis that MAEL HMG-box is a unique, RNA-binding domain.

ACKNOWLEDGEMENTS

I am grateful and thank Alex Bortvin[1] for mentoring me scientifically and personally through my graduate career in his lab.

I thank the following for contributing technical expertise, materials, and resources for this work: Rejeanne Juste[1] for performing cloning and mutagenesis; Frederick Tan[1] for help with computational analysis; Johns Hopkins Center for Molecular Biophysics for the acquisition of CD measurements; Xin Chen[2] for providing fly testis cDNA; Jon R. Lorsch[3] for advice with PC column purification and Joseph-Kevin Igwe for performing HMGB1a purification.

I thank the following people for critique and discussion of the dissertation work: Safia Malki[1], Valeryia Gaysinskaya[1,2], Julio Castañeda[1,2], Chen-Ming Fan[1]

I thank the past and present members of my thesis advisory committee for their support, guidance, discussions and suggestions: Alan Spradling[1], David Zappulla[2], Sarah Woodson[2] and Evangelos Moudrianakis[2].

I thank Christoph Lepper[1] for his critique, discussion and help with thesis review.

[1] Carnegie Institution of Washington Embryology

[2] Johns Hopkins University

[3] National Institute of General Medical Sciences

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

ABREVIATIONS

ss – single-stranded

ds – double-stranded

TE – transposable element

piRNA – piwi-interacting RNA

fruit fly – *Drosophila melanogaster*

HMG – high mobility group

IEM – immune electron microscopy

PGC – primordial germ cells

LINE-1/ L1 – long interspersed nuclear element 1

SINE – short interspersed nuclear element

TPRT – target-primed reverse transcription

4WJ – four-way junction

SS – sequence-specific nucleic acid binding

NSS – non-sequence-specific nucleic acid binding

SRY – sex determining region of Y

HMGB1a – high mobility group box 1 protein box A

Mael – *Drosophila melanogaste*r Maelstrom protein

MAEL – *Mus musculus* Maelstrom protein

MSD – Maelstrom-specific domain

PTM – post-translational modification

CHAPTER I

INTRODUCTION

INTRODUCTION – PART I

piRNAs, transposon silencing, and germline genome integrity

Reprinted from Mutation Research/Fundamental and

Molecular Mechanisms of Mutagenesis in accordance with Elsevier author rights

piRNAs, transposon silencing, and germline genome integrity

Julio Castañeda*

Pavol Genzor*

Alex Bortvin

*These authors contributed equally to this work

Biology Department, Johns Hopkins University, Baltimore, MD 21218

Department of Embryology, Carnegie Institution for Science, Baltimore, MD 21218

Address correspondence to: Alex Bortvin, email: bortvin@ciwemb.edu, 3520 San Martin Drive, Baltimore, Maryland 21218, Tel: (410) 246-3034, Fax: (410) 243-6311

Keywords: piRNA, transposable elements, TE, LINE-1, epigenetic, DNA methylation, histone modifications, germline, mouse, mechanism, Dnmt, Maelstrom

Abstract

Integrity of the germline genome is essential for the production of viable gametes and successful reproduction. In mammals, the generation of gametes involves extensive epigenetic changes (DNA methylation and histone modification) in conjunction with changes in chromosome structure to ensure flawless progression through meiotic recombination and packaging of the genome into mature gametes. Although epigenetic reprogramming is essential for mammalian reproduction, reprogramming also provides a permissive window for exploitation by transposable elements (TEs), autonomously replicating endogenous elements. Expression and propagation of TEs during the reprogramming period can result in insertional mutagenesis that compromises genome integrity leading to reproductive problems and sporadic inherited diseases in offspring. Recent work has identified the germ cell associated PIWI Interacting RNA (piRNA) pathway in conjunction with the DNA methylation and histone modification machinery in silencing TEs. In this review we will highlight these recent advances in piRNA mediated regulation of TEs in the mouse germline, as well as mention the repercussions of failure to properly regulate TEs.

Mammalian Germline Specification and Reprogramming

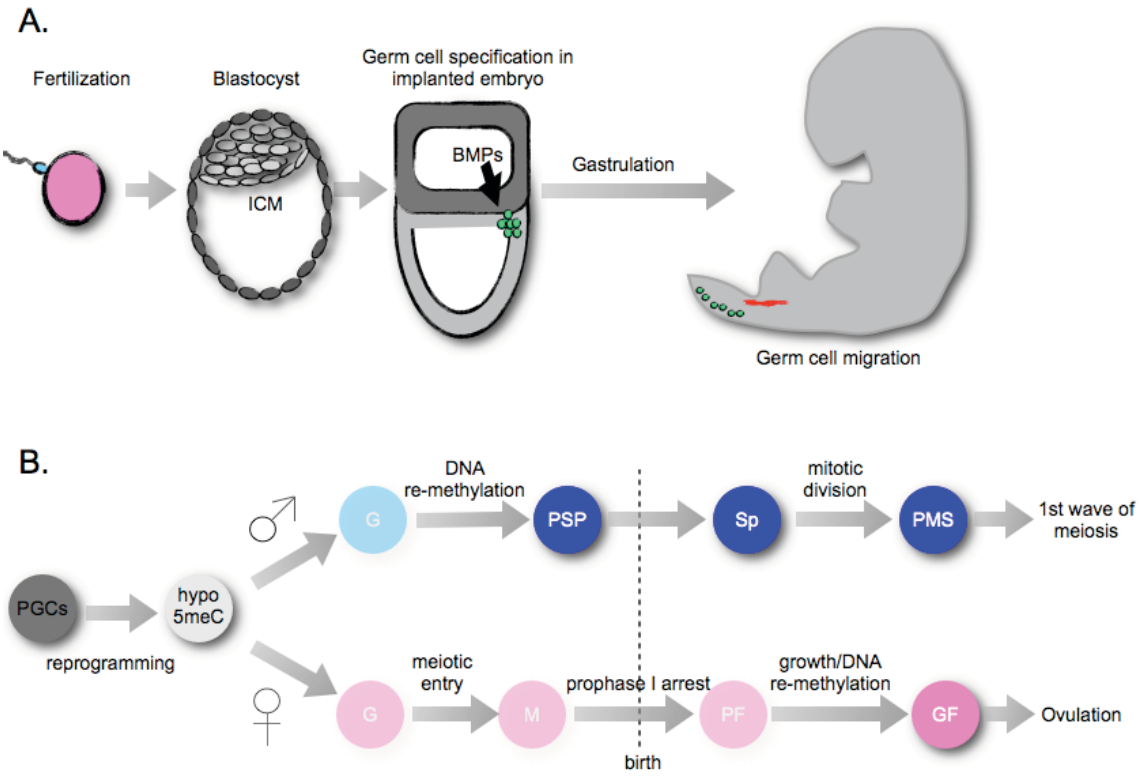
Following fertilization, male and female haploid pronuclei fuse to form the diploid nucleus of the mammalian zygote. After several rounds of cleavage, the blastocyst is generated (at embryonic day 3.5, E3.5) where a group of interior cells, called the inner cell mass, will give rise to the embryo proper. The identity of these early embryonic cells is thought to be determined and established based on their position and inductive cues

they receive from neighboring cells. Prior to gastrulation (E6.5), a small subset of cells at the junction between the embryo and extra-embryonic tissue (the posterior epiblast) is induced by ligands of the Bone Morphogenic Protein (BMP) family to initiate a germ-cell developmental cascade (Figure 1A) [1-4]. After gastrulation, these germ cell precursors (often referred to as primordial germ cells or PGCs) reside at the most posterior region of the epiblast and then migrate to the genital ridge, the future gonad. During germ cell migration, PGCs begin to undergo epigenetic changes that include loss of transcriptionally repressive marks such as DNA methyl-cytosine [5, 6]. When germ cells reach and colonize the genital ridge (E10.5), the rate of epigenetic changes increases rapidly and in the case of DNA de-methylation is completed within days [7]. This process (called "epigenetic reprogramming" or simply "reprogramming") is thought to allow PGCs to reacquire a pluripotent state [8] and to establish a "blank" genome for which to "paint" the sex-specific imprints of the embryo [9-13].

Unlike this equal epigenetic erasure that occurs contemporaneously in both sexes [14, 15], the timing of meiotic entry and the re-establishment of DNA methylation between males and females is dimorphic (Figure 1B) [16]. In females, the entire germ cell pool enters meiosis before DNA re-methylation during embryonic development (E13.5), arrests at prophase I, and establishes primordial follicles. At puberty and at each subsequent estrous cycle, a subset of primordial follicles are recruited for maturation where growth and re-methylation occurs [17]. The completion of the first round of meiosis (MI) occurs during ovulation while the second round (MII) is only completed upon fertilization of the ovum.

Figure 1. Germline Specification and Reprogramming:

A) After fertilization, cellular division produces the blastocyst containing the inner cell mass (ICM - light grey). After implantation, the ICM will give rise to the epiblast (light grey) from which the embryo proper forms. BMP signals from extra-embryonic tissue (dark-grey) induce primordial germ cells (PGCs - green). After gastrulation, PGCs end up at the posterior of the embryo and migrate to the genital ridge (red), (extra-embryonic tissue not shown).B) Epigenetic reprogramming and meiotic entry in male and female germ cells. PGCs in the gonad undergo DNA demethylation (indicated by lighter shading). Male germ cells re-establish DNA methylation (dark blue) before meiotic entry after birth. Female germ cells re-establish DNA methylation (dark pink) after birth. PGCs - primordial germ cells, hypo 5meC - de-methylated PGCs, G - gonocyte, PSP - prospermatogonia, Sp - Spermatogonia, PMS - pre-meiotic S-phase, M - meiocyte, PF - primordial follicle, GF - growing follicle.


A.
Fertilization
Blastocyst
Germ cell specification in implanted embryo
ICM
BMPs
Gastrulation
Germ cell migration
B.
DNA re-methylation
mitotic division
PGCs
hypo 5meC
reprogramming
G
PSP
Sp
PMS
1st wave of meiosis
meiotic entry
prophase I arrest
growth/DNA re-methylation
G
M
PF
GF
Ovulation
birth

In males, DNA methylation in germ cells (termed prospermatogonia at this time) begins to be reestablished during embryonic development (E18.5) and is completed by postnatal day 2 (P2). Meiosis in males initiates from spermatogonia, the stem cell pool established by prospermatogonia, after birth in waves spaced several days apart (10 days in the mouse) and is completed within one to two weeks, depending on the species (~12 days in the mouse) [18]. The accurate progression through meiosis in males is highly contingent on the re-establishment of methyl-cytosine marks that were erased in PGCs [19-21]. Female meiosis is also dependent on DNA methylation [22]; however, meiosis in females can be completed in the absence of the DNA methylation machinery required for male meiosis [19-21]. Thus, DNA methylation is essential to the mammalian germline in both sexes, however, to varying degrees.

A major reason, although not exclusive, for the re-establishment of DNA methylation is that it is the primary mode of silencing TEs in mammals. TEs are selfish DNA elements highly abundant in mammalian genomes, which cause DNA damage via transposition [23-25]. Deposition and maintenance of methyl-cytosine is mediated by the DNA methyltransferases (DNMT), which are required to silence expression of TEs [26]. In addition to the DNMTs, work within the last several years has shown that the piRNA pathway is essential for *de novo* DNA methylation of TEs [27, 28]. Members of this pathway associate with germline enriched piRNAs (26-31nts long), allowing for recognition and silencing of TEs. Why is epigenetic silencing of TEs, especially in the context of the germline, so important in mammalian germline development? What are the key pathways (their constituents and molecular mechanisms) involved in silencing of TEs, and what are the repercussions of their absence? These and other questions will be

discussed in the following sections.

Eukaryotic Genomes and the Transposable Element Challenge

Metazoan genomes are generally large, containing DNA measuring in the thousands of megabases; however, the fraction that had been thought to be functionally important (coding sequences and their regulatory elements) comprises less than two percent of the genome depending on the organism [24, 29]. The Human Genome Project revealed many surprises about the organization and content of the human genome, one of which is the large presence of repetitive elements. Repetitive elements within the human genome include transposable elements (TEs), mobile pieces of DNA that were initially discovered in maize by Barbara McClintock [30]. TEs make up roughly ~45% of the human genome and can be divided into DNA and RNA TEs [24]. DNA based transposons propagate by a "cut and paste" mechanism using a transposase enzyme for their excision and insertion [31]. RNA TEs replicate via an RNA intermediate and comprise up to 42% of the human genome. These TEs consist of Long Terminal Repeat (LTR) and non-LTR TEs. Structurally, LTR TEs resemble and appear to be forbearers of retroviruses that gained the capacity for horizontal transfer through acquisition of an *Envelope* gene [32]. The non-LTRs TEs represent the biologically most relevant class since they are the only active transposons in humans and comprise the majority of TEs in the human genome (~34%) [24, 29, 33, 34]. This class can be subdivided into autonomous (TEs capable of transposition) and non-autonomous (TEs dependent on autonomous elements for transposition) that comprise 21% and 13% of the genome, respectively. Long interspersed nuclear elements (LINE-1 or L1) are the autonomous elements, while the

non-autonomous include short interspersed nuclear elements (SINEs). Both autonomous and non-autonomous transposons are of serious concern for germline development as indicated by the activation of these elements during the reprogramming window. The spotlight belongs especially to L1 elements, whose ability to retrotranspose themselves and mobilize other non-autonomous elements has been linked to reproductive disorders and other diseases [35-37].

LINE-1 Transposable Elements as Mutagens

L1 expression can be detected in the germ line, during embryonic development, in neuronal tissue, and cell lines derived from various cancers [38-44]. The L1 life cycle begins with transcription of the element by RNA polymerase II. The L1 mRNA encodes two proteins [Open Reading Frame 1 and 2 proteins (ORF1p and ORF2p) described in the following section] that facilitate L1 mRNA reverse transcription and integration at novel locations in the genome (Figure 2A). Detailed mechanisms of L1 insertion into the genome (termed target-primed reverse transcription, TPRT) have been reviewed previously [45-47], however, it should be emphasized that the complete mechanism of L1 insertion is not fully understood. Insertion into new genomic locations is detrimental to genome integrity as it produces DNA breaks and has the potential to disrupt gene-coding regions [48]. For example, L1 mediated mutagenesis described 20 years ago, was shown to be the causative agent of hemophilia A [49]. In addition, an increase association of L1 and its encoded proteins in human diseases (i.e. cancers) [37, 50] suggests TEs as the etiological basis for these diseases and mandates a better understanding of proteins encoded by L1 elements and of their regulation.

ORF1p and ORF2p

The first open reading frame of the L1 dicistronic mRNA encodes ORF1p [51]. ORF1p is a roughly 40kD RNA binding protein that forms a trimer in solution and binds L1 mRNA via its noncanonical RNA-recognition motif [52-58]. ORF1p preferentially binds its coding mRNA thereby facilitating cis-TPRT (Figure 2A) [54, 57, 59]. The binding of ORF1p to L1 RNA is independent of ORF2p and is important for the formation of ribonucleoprotein complexes thought to chaperone L1 mRNA back into the nucleus [60-62]. The second open reading frame, ORF2p, encodes a 150kD protein with endonuclease (EN) and reverse transcriptase (RT) activities [63, 64]. Both ORF1p and ORF2p are required for retrotransposition of L1. Besides facilitating L1 transposition, ORF2p is required for non-L1 element transposition such as SVA or SINEs in trans by a similar TPRT mechanism [65]. Retrotransposition assays in *HeLa* cells have shown that EN and RT activities of ORF2p alone are sufficient for SINE retrotransposition, but the efficiency of this process is increased in the presence of ORF1p [66]. Expression of L1 mRNA and its encoded proteins poses significant threats to the genome. The ramifications are most serious in the context of the germline, since resulting defects can be fixed and inherited by successive generations thereby expanding the active TE population. In order to keep the TE threat at bay, epigenetic control mechanisms, described in more detail below, have been adapted and refined over evolutionary time to ensure the silencing of these parasites.

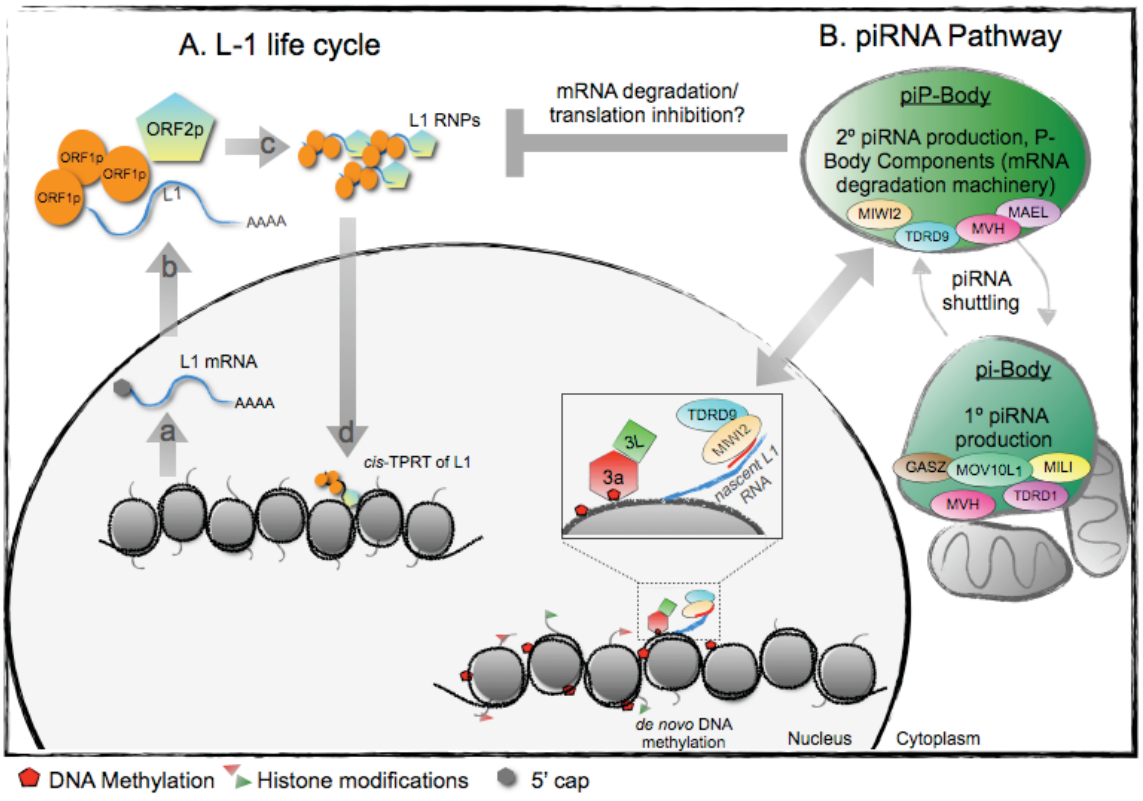


Figure 2. Transposon regulation in the germline: A) Transposable elements are transcribed during the period of genome reprogramming (a). TE mRNA is then transported to the cytoplasm (b), encoded proteins (ORF1p and ORF2p) are translated allowing for the assembly of the ribonucleoprotein particles (RNPs) (c). RNPs are then transported back into the nucleus and integrated into new genomic locations (d) via target-primed reverse transcription (TPRT), which consequently results in genomic instability. B) Primary (1º) and secondary (2º) piRNAs are proposed to be generated in the pi-Body and piP-Body, respectively. These piRNAs facilitate *de novo* DNA methylation (by DNMT3a/DNMT3L complex) to silence transposon expression. The exact mechanism of piRNA-guided DNA methylation remains to be elucidated, however, several implicated genes in this pathway are shown.

Epigenetic Reprograming

DNA methylation and L1 expression

Erasure of methyl-cytosine marks during germ cell development (starting at E10.5 in mouse PGCs) is crucial for establishment of sex-specific imprinting [26], but also provides an opportunistic environment for L1 propagation as methylation of the upstream regions of L1 is essential to maintain their silenced state [67-69]. In the male germline, this mark is re-established by the *de novo* methyltransferases DNMT3a and 3L and maintained by DNMT1 [20, 21, 70-75]. Deletion of DNMT3a or 3L results in activation of TEs in the male germline resulting in apoptosis of spermatocytes and sterility. Interestingly, in male germ cells lacking DNMT3a or 3L the methylation patterns of satellite DNA remain unperturbed suggesting this complex specifically targets TEs [20, 74]. The female germline can develop mature oocytes and ova without DNMT3a, 3b, or 3L but the resulting ova are incapable of supporting embryonic development [19-21]. Mechanistically, DNMT3L binds to and stimulates DNMT3a, which then transfers a methyl group from S-adenosyl-L-methionine to the fifth carbon on cytosine [75]. The DNA substrate preference for this complex has been suggested to be transposon regions enriched in CG dinucleotides [76, 77], however, the exact mechanism of how the DNMT3 complex is targeted to specific regions of the genome (i.e. TEs) is not clear. One possibility is that the DNMT3 complex is guided to genomic sites through complementarity to a small RNA as is the case with *Arabidopsis* [78, 79]; however, Ross et al. have called this idea into question using an *in vitro* approach [80]. It has been shown that DNMT3L interacts with the non-methylated tail of histone 3 (H3) [81, 82]. The presence of H3 lysine 4 methylation (H3K4) abolishes this interaction and prevents

de novo DNA methylation [81, 82]. These results suggest H3K4 methylation could be the targeting mechanism for the DNMTs to regulate to TEs.

Histone modifications and TE silencing

Recent work has shown other histone tail modifications involved in regulating TEs in addition to H3K4 methylation. Acetylation of histone tails was implicated in human embryonic carcinoma cell lines where the silencing of a L1 transgene was alleviated by the addition of the histone deacetylase inhibitors [83]. Biotinylation of H4K12 and H2AK9 was shown to be essential for silencing TEs in human and mouse cell lines and *Drosophila* [84, 85]. In addition, increased levels in acetylated H3 and H3S10 phosphorylation upregulate expression of specific transposable elements [86, 87]. Work in *Arabidopsis* implicated deubiquitination of H2B and H3K9 demethylation with TE expression [88]. These results lead to further questions about whether there is one or a combination of histone modifications specific to TEs. Indeed, on a genome-wide level 38 histone modifications show enrichment at TE regions, further suggesting more complex regulation besides H3K4 methylation [67]. However, detailed analysis of histone modifications of TEs in animal models have not been performed thus far. The modification of histone tails could be the targeting mechanism for the DNMT3 enzyme complex, however, this begets the question as to how the histone modifiers themselves are targeted to TEs within the genome. Recent work, which we will describe below, has shown that the additional targeting mechanism for histone modifiers could include the piRNA pathway, a class of small RNAs predominately expressed in the germline, thought to facilitate genomic TE recognition.

The piRNA Pathway-Mechanistic Overview

The Discovery of piRNAs in Drosophila and Mouse

The existence of piRNAs were hinted at in *Drosophila* [89, 90] but were named repeat-associated RNAs (rasiRNAs) [91]. The breakthrough in understanding piRNA biology, however, was not until next generation sequencing became available. In 2006, four groups reported the discovery of RNAs bound to the rodent homologs of PIWI proteins [92-95], demonstrating that these were longer (26-31nt) than miRNAs and siRNAs, were encoded in clusters throughout the genome, predominately correspond to TE sequences, and are specific to testes. All four studies differentiated these novel small RNAs (named piRNAs for their association with PIWI proteins) from miRNAs and siRNAs since piRNAs mapped to regions that did not produce either a dsRNA or hairpin intermediate and appeared to originate from transcripts several kilobases long [92-95]. Shortly thereafter, *Drosophila* rasiRNAs from testes were characterized as piRNAs and shown to be generated by a distinct pathway from miRNAs and siRNAs [96].

The Primary piRNA Pathway

Characterization of piRNAs since 2006 has shown that piRNAs are generated via two distinct molecular mechanisms, the primary and secondary/amplification (or "ping-pong") pathways [97-99]. Considering the large tandem arrays of piRNAs in the genome, primary piRNAs are thought to be transcribed as long ssRNA transcripts [28, 97]. However, whether a multi-kb RNA transcript is produced and gets processed into mature piRNAs remains to be determined. The molecular players at the top of the primary pathway have only recently been described using *Drosophila* genetics. Four groups

independently showed that an RNA helicase (Armitage, Armi), a Tudor domain protein (Yb), and an exonuclease (Zucchini, Zuc) are required for the generation of primary piRNAs that are loaded onto Piwi protein [100-104]. A mechanistic model for primary piRNA production based on these papers suggests that Yb may act as a scaffold (analogous to murine Tudor domain proteins) bringing together all the players into a cytoplasmic body (the Yb body) [105]. Armi and Zuc, in the Yb-body, process pre-piRNA transcripts into mature piRNAs that are then loaded onto Piwi, allowing Piwi to enter the nucleus and mediate silencing [100-102, 104]. However, the molecular mechanism of downstream TE silencing in *Drosophila* is unclear. It is possible that piRNAs recognize RNA transcripts and target them for degradation since the Yb body sits adjacent to Processing Bodies (sites of RNA degradation) [105, 106]. Alternatively, silencing may include piRNA-mediated recognition of DNA targets that then mediate epigenetic changes of these genomic targets to keep them off.

The Secondary/"Ping-pong" Pathway

The main effectors of the "ping-pong" pathway are the two other *Drosophila* encoded PIWI proteins, Aubergine (Aub) and Argonaute3 (Ago3). Initiation of piRNA mediated silencing begins with transcripts from piRNA loci (antisense with regard to TEs), that are processed to functional size RNAs and bound by Aub [97]. The Aub-piRNA complex directs the cleavage of sense TE transcripts to generate degradation products that are processed and loaded onto Ago3 [97]. The Ago3-piRNA complex can then direct the cleavage of antisense transcripts to produce sense degradations products that also get processed and loaded onto Aub, thus completing the loop, or so called "ping-pong" cycle.

The main signature of the secondary pathway is the complementarity between the first 10nt of Aub and Ago3 piRNAs, where the uridine (U) predominates as a first base of Aub piRNAs, and adenine (A) as tenth of Ago3 piRNAs [97]. This 1U-10A "ping-pong" signature is absent from the primary piRNA pathway in *Drosophila* [98, 99, 107]. As a result of back-and-forth interactions of Aub and Ago3, genomic or maternally deposited piRNAs can be amplified and their targets silenced [97, 108]. The downstream mechanism TE silencing by the piRNA pathway in *Drosophila* is thought to include recruitment of the RNA degradation machinery and a yet to be described epigenetic silencing mechanism that keeps TEs in a silent state [109, 110]. In the case of the mouse piRNA pathway, the major downstream epigenetic mechanism implicated in repression of TEs is the establishment of DNA methylation, which is absent in *Drosophila*. Major murine counterparts of the *Drosophila* piRNA pathway components are described in detail below.

The piRNA pathway in the Male Mouse Germline

When mouse PGCs enter the genital ridge (E10.5), DNA methyl marks on cytosine are erased [7], relieving the suppression of TEs. The especially prominent expression of L1 TEs is corroborated by the presence of ORF1p at E15.5 and E14.5 in females and males, respectively [111, 112]. It is essential that DNA methylation of these TEs be re-established in embryonic male germ cells; failure to do so results in TE transcript accumulation, transposition-induced DNA damage, defects in homologous chromosome synapsis, meiotic arrest, and sterility [20, 74, 113]. Interestingly, in the male, certain piRNA pathway components [the PIWI proteins MILI & MIWI2, Maelstrom (MAEL),

GASZ, Mouse Vasa Homolog (MVH), TDRD1, and TDRD9] begin to accumulate before ORF1p is detected as if anticipating the release of L1 inhibition [28, 112, 114-116]. Unlike in *Drosophila*, there is a substantial presence of both primary and "ping-pong" cycles in male mouse embryonic germ cells which complicates the molecular dissection of this pathway [28, 112]; nevertheless, many studies of mouse mutants have elucidated certain aspects of the piRNA pathway in the male germline (Table 1).

The pi-Body: the presumed site of the Primary piRNA pathway

The upstream components of the primary pathway include MOV10L1, GASZ, MILI, TDRD1, and MAEL. The cytoplasmic organization of the male embryonic pathway is such that most of the upstream components (MOV10L1, GASZ, MILI, and TDRD1) reside in the piRNA-Body or pi-Body (previously known as the intermitochondrial cement) while MAEL (along with MIWI2, and TDRD9) resides in the adjacent cytoplasmic structure termed the piRNA-Processing Body or piP-body (Figure 2B) [112, 115, 116]. Of these primary components, only MAEL has been localized to the nucleus by immunofluorescence [112]. MOV10L1 (a homolog of Armi) is required for the generation of primary piRNAs, which in the mouse are sense with respect to TEs [117, 118]. In *Mov10l1* mutants, total small RNA deep sequencing at P10 fails to detect piRNAs and immunoprecipitation of MIWI2 and MILI at P0 (before meiosis initiates) shows that these proteins are devoid of piRNAs as well [118]. MOV10L1, like *Drosophila* Armi, is therefore required for the primary piRNA pathway. piRNA defects in mutants for the mouse homologues of Yb (TDRD12) and Zuc (PLD6) are yet to be reported. GASZ, a protein with many protein-protein interaction motifs [119], was found

to be required for correct localization of MILI and for expression of many piRNA pathway proteins [115]. The mislocalization of MILI in *Gasz* mutants suggests GASZ may play the role of *Drosophila* Yb (which localizes PIWI) in the mouse possibly with TDRD12 (Yb). In the *Gasz* mutant, piRNAs are drastically reduced but not absent at P7, P10, and P14 based on total small RNA sequencing. Germ cells deficient for TDRD1 show a disproportionate accumulation of sense, exonic-derived piRNAs (as opposed to piRNAs from non-coding regions such as TEs) in purified MILI complexes by P15, which suggests that TDRD1 helps direct proper piRNA selection and loading onto MILI [120, 121]. However, the localization of MILI to the pi-body is not affected in *Tdrd1* mutants.

Mael mutants show an interesting phenotype in embryonic and perinatal male germ cells. At E16.5, MAEL deficient germ cells correctly localize MILI and TDRD1, however they show a complete absence of piRNAs suggesting that MAEL functions in the primary piRNA processing pathway unlike *Drosophila* Mael [100, 112, 113]. Perhaps MAEL aids in the initial production of piRNAs as it is the only primary piRNA component localized to the nucleus, however, why MAEL also localizes to a distinct cytoplasmic body from the other primary components is unexplained. Unexpectedly in the *Mael* mutant, at P2 piRNA profiles recover to near wild-type levels. piRNA recovery may result from spurious degradation products generated by the RNA degradation machinery leading to the production of small RNAs that encounter MILI, which then initiates the “ping-pong” cycle [122]. Since total small RNA profiles from embryonic testes in *Mov10l1*, *Gasz*, and *Tdrd1* mutants were not analyzed, this piRNA recovery seen in *Mael* mutants may not be unique.

Absence of MILI results in a loss of most piRNAs corresponding to TEs [27, 28]. The loss of these piRNAs leads to a failure to re-methylate TEs in male embryonic germ cells, which then leads to massive upregulation of L1 in meiosis. Taken together, MOV10L1 and MAEL appear to be responsible for the production of piRNAs, GASZ for the correct localization of MILI, and TDRD1 for proper piRNA loading of MILI, which is essential as MILI is the main effector protein of the pathway. Similar to the *Mili* phenotype, absence of MOV10L1, GASZ, and TDRD1 results in activation of L1 during male meiosis as a consequence of their unsuccessful DNA methylation, which results in DNA damage, synapsis and pairing defects, meiotic arrest, apoptosis, and ultimately sterility [27, 28, 112, 113, 115, 117, 118, 123].

The piP-Body: the presumed site of the Secondary piRNA pathway

Major players of secondary piRNA pathway are piP-body resident proteins - MIWI2, TDRD9, and MVH (Figure 2B) [112, 116]. The antisense identity of piRNAs associated with MIWI2 indicate that MIWI2 is downstream in the ping-pong cycle, associating with piRNAs generated by MILI presumably in the pi-body [28]. Complementarity of MIWI2 antisense piRNAs with L1 transcripts suggests an mRNA recognition and degradation pathway involving MIWI2's PIWI domain [27, 116, 124-126]. A direct role of MIWI2 mediated cleavage as the sole mechanism of TE transcript degradation is complicated by the fact that Processing Body (P-body) components also localize with MIWI2 [112]. P-bodies have been shown in yeast and cell culture studies to be sites of mRNA degradation and mRNA storage [127, 128]. It is possible that MIWI2 recruits P-Bodies to aid in the degradation of TE transcripts; however, MIWI2 can also localize to the nucleus where

piRNAs associated with MIWI2 may recognize TE nascent transcript by complementarity that then recruit the DNMT proteins [27, 28]. TDRD9, which is the mouse homolog of Spn-E^{Hls}, also has a similar localization pattern as MIWI2 suggesting that it aids MIWI2 in generating secondary piRNAs and helping in an RNA directed DNA methylation (RdDM) process [116, 129-132]. Indeed, *Tdrd9* mutants show a decrease in secondary piRNAs by total small RNA deep sequencing, however, the localization of MIWI2 is unaffected in the absence of TDRD9 [116]. The correct localization of MIWI2 in the *Tdrd9* null background suggest that either MIWI2 can localize to the nucleus on its own or that there is another component that aids in its localization.

MVH (an RNA helicase) localizes to both the pi-body and the piP-body [133, 134]. In the *Mvh* mutant, embryonic germ cells show a decrease (20% of normal levels) but not an absence of piRNAs, with a preferential loss of MIWI2 associated piRNAs. In fact, immunoprecipitation of MIWI2 fails to detect associated piRNAs, suggesting MVH facilitates their loading onto MIWI2. Loss of MIWI2 piRNAs and sequence analysis of the remaining piRNAs suggest that the primary pathway is intact but the secondary pathway is compromised. The localization of the primary and secondary pathway components (MIWI2 and TDRD9) in embryonic germ cells shows that only TDRD9 is correctly localized in the absence of MVH, which is at odds with studies implicating other upstream components in TDRD9 localization [116, 134] MILI and TDRD1 mislocalization and concurrent production of primary piRNAs (albeit at a lower level to wild type) suggests that the primary pathway does not require proper pi-body formation. The inability to load MIWI2 with piRNAs, however, suggests that pi-body formation

and/or piP-body localization of MIWI2 is required for piRNA loading of MIWI2. Since MVH is an RNA helicase and localizes to both the pi-body and the piP-body, it is tempting to speculate that MVH shuttles piRNAs between these bodies, however, MVH does not associate with piRNAs [134].

Mutations in all the above-mentioned mammalian piRNA components (MILI, MIWI2, MAEL, MOV10L1, GASZ, MVH, TDRD1, and TDRD9) lead to a reduction of piRNAs and failure to silence TEs (Table 1). Although all the phenotypes suggest an RdDM mechanism, a direct association between several of the above mentioned components (in particular MIWI2) fail to detect the presence of any DNMT proteins (either 1 or 3), and the reverse is also true (DNMT3L immunoprecipitation fails to detect piRNA proteins) [81, 135, 136]. These data suggest there is an intermediate step between MIWI2 recognition of TEs and recruitment of DNMTs. It seems likely that the gap between TE methylation and piRNA mediated silencing is bridged by modification of histone tails, which could function to recruit the DNMTs to specific sites. However this speculation has not yet been corroborated by experiments.

Summary and Perspectives

Proper repression of TE activity is vital for maintaining the integrity of the genome. In order for the development of gametes competent for fertilization TEs must be silenced. In the male germline, silencing of TEs (in particular L1) occurs before meiosis, and is achieved by the DNMT enzymes in conjunction with the piRNA pathway.

Mouse Gene	Drosophila Homolog	Protein Domains	Function	Mouse Mutant Phenotype	Reference
Mili	Aubergine	PAZ, PIWI	Binds 1º piRNAs, generates 2º piRNAs, required for piP-Body formation	Loss of most piRNAs, failure to methylate TEs, upregulation of TEs at meiosis	26,27, 87,108, 114
Miwi2	Ago3	PAZ, PIWI	Binds 2º piRNAs, possible effector of RNA-directed DNA methylation	Loss of 2º piRNAs, failure to methylate TEs, upregulation of TEs at meiosis	26,27,108, 115
Mov10 L1	Armitage	DEXDc-helicase, AAA ATpase	Generates 1º piRNAs and/or loads piRNAs onto Mili	Loss of all piRNAs, failure to methylate TEs, upregulation of TEs at meiosis	108,109
MVH	Vasa	DEADc-helicase, HELICc	Required for "ping-pong" pathway & loading Miwi2 with piRNAs	Reduced levels of piRNAs (20% of WT), mislocalization of many piRNA components, upregulation of TEs at meiosis	124,125
Tdrd1	CG14303	MYND domain, 4 Tudor domains	Ensures proper piRNA loading onto Mili, required for piP-Body formation	Increase in exon-derived piRNAs onto Mili, failure to methylate TEs, upregulation of TEs at meiosis	111,112
Tdrd9	Spn-EHls	DEXDc-helicase, HELICc, HA2, 1 Tudor domain	Ensures production of 2º piRNAs	Reduction of 2º piRNAs, increase of 1º piRNAs, failure to methylate TEs, upregulation of L1 at meiosis	107
Tdrd12	Yb	1 Tudor domain	Not examined	None reported	N/A
Pld6	Zucchini	PLDc	Not examined	None reported	N/A
Mael	Mael	HMG-like Box	Required for proper timing of 1º piRNA production, required for piP-Body formation	Loss of all piRNAs at E16.5, recovery of piRNAs by P2, failure to maintain TE methylation at meiosis	104,105
GASZ	CG2183?	4 Ankaryin Repeats, Sterile Alpha Motif	Required for pi-Body formation and expression of piRNA components, possibly facilitates protein-protein interactions	Reduction of 1º and 2º piRNAs, mislocalization of Mili & Tdrd1, reduced expression of piRNA pathway components	106,110

Table 1. Mouse embryonic piRNA pathway components: Genes implicated in the mouse embryonic piRNA pathway, their *Drosophila* homologs, encoded domains, and deduced function based on mutant phenotype. 1º - primary, 2º - secondary

Failure to silence these elements in the male leads to meiotic arrest and apoptosis due to massive DNA damage cause by unchecked L1 insertion events. In addition to the germline, L1 activity has the potential to compromise genome integrity in all cells as each cell contains thousand of copies of L1 encoded in their genome. Indeed, L1 activity has been linked to many cancers and correlated with several diseases [36]. The question remains whether the association of L1 TEs with these cancers is a causative agent or just a correlation. If this is just correlated, L1 may serve as a useful diagnostic tool. For example, studies in breast cancers have correlated nuclear localization of ORF1p with decreased survival of patients compared to patients with predominantly cytoplasmic ORF1p [38].

While the activity of TEs is largely viewed to have a negative consequence for genome integrity, there are several lines of evidence that suggest transposons and transposon encoded genes exert a positive effect. In one case, transposition of the Het-A DNA transposon comprises and maintains the telomeres of *Drosophila* [137, 138]. In several other cases, it has been the usurpation of TE encoding genes for the benefit of the host. The most well known example in humans are the RAG recombinases, which are thought to have originated from a DNA transposon [139]. A recent study of a ciliated protozoan, *Oxytricha trifallax*, demonstrated that TE derived transposases facilitate large-scale genome rearrangements and removal of transposon sequences in the maturing macronucleus [140-142]. Mammalian genome-wide analysis suggests that TEs influence mutation and recombination rates in the germ line [143]. In addition, the TE landscape in various genomes may be indicative of their role in chromosome biology and evolution, for example there is an enrichment of TE elements on sex chromosomes [144].

Furthermore, multiple reports implicate endogenous reverse transcriptase (ORF2p of L1) to be involved in regulation of the embryonic gene activation, modulation of growth, and proliferation of germ cells [145-147]. However, more in-depth understanding of TEs, their encoded proteins, and their effects within their genomic location is required to be able to attribute specific functions to transposons.

TEs have also been useful for the biomedical community. The ability to randomly integrate into the genome has made TEs a prime subject for developing genetic tools [148]. One of the most well established system is the P-element (a DNA transposon) used to manipulate the *Drosophila* genome [149, 150]. The P-element has been utilized in numerous genetics screens and has lead to production of thousands of genetic lines that have been invaluable in dissecting multiple pathways in numerous tissues [151]. In fact, a P-element genetic screen uncovered PIWI (P-element Induced WImpy testis) [152]. A more recent and versatile TE is the *Sleeping Beauty (SB)* transposon, a "resurrected" transposon capable of hopping in numerous genomes [153]. Mobilization of this element has been shown to result in somatic and embryonic cancer when induced [154-157]. The SB TE has also been utilized to generate transgenic rats as well as in gene therapy using animal models [158]. Modifications of TEs to carry vital genes and to control their insertion will be of great interest for future gene therapy trials.

Although much information has been gained from work on TEs and the piRNA pathway, many questions remain. It is beyond the scope of this review to list all the questions here, however, we would like to highlight some of the most intriguing avenues that are currently being pursued. In the immediate future, the field of mammalian piRNAs would be advanced by an understanding of the mechanism that ties the piRNA

pathway with methylation of DNA, as neither PIWI proteins nor DNMTs associate together in a complex [81, 135, 136]. Concerning the female piRNA pathway, very little is know as most piRNA pathway mutants are female fertile [113, 115, 116, 123, 133, 159]. Does the lack of a female piRNA phenotype suggest there is no role for piRNAs in the female germline and/or that there is a redundant mechanism that ensures the silencing of TEs in primordial follicles [160]? Perhaps the difference between the male and female germline in their dependency on the piRNA pathway may be attributed to the difference in timing of DNA methylation in meiosis. The post-meiotic methylation of the female genomes leads us to question how the primordial follicles are protected from TEs during meiosis. On a global scale, should there be a concern about the increasing use of RT inhibitors used to combat the global AIDS pandemic considering the endogenous L1 activity in sperm, embryos and in neurons [38-44, 145-147]?

We have come a long way in understanding the biology of transposable elements and the mechanisms that regulate them since the first discovery of these elements in 1950. With the increasing availability of next generation sequencing, more sophisticated whole-animal genetic manipulations, and advancing live-cell imaging techniques we are on the verge of some remarkable discoveries regarding TEs, their role within our genome, and the potential they hold for our evolution.

INTRODUCTION – PART II

Maelstrom

The Maelstrom gene was identified via a P-element-induced female-sterile screen in *Drosophila melanogaster* where its deletion led to mislocalization of multiple RNAs involved in early-oocyte development and axis establishment. In this first study, Mael was suggested to play a role in anchoring MTOC in early oocyte and/or short-range transport of RNAs to proper sites [161]. Later Mael was shown to co-localize with Vasa, an RNA helicase, in cytoplasmic structures found in the periphery of nuclei of germ cells known as nuages [162-164]. In addition to nuage, in the fruit fly, Mael shuttled between nuclear and cytoplasmic compartments [165], and chromatin immunoprecipitation (ChIP) of fruit fly Mael suggested that it may be binding DNA and, by doing so, regulated germ cell fate [166]. In the mouse, MAEL was shown in proximity to nuclear pores, in cytoplasm, but unlike in the fruit fly, it was enriched in the cytoplasmic nuages [112]. Despite its low nuclear abundance in the mouse, the results from the fruit fly suggested that Mael might be a DNA-binding protein influencing chromatin.

piRNAs On The Scene

Meanwhile, a novel type of small RNAs that interact with piwi proteins (piRNAs) appeared on the scene as prominent regulator of germ-cell development. piRNAs are 26-31 nt in length and originate from genomic positions that either correspond to transposable elements or to the UTR regions of genes. These mRNAs sediment in heavy polysome fractions; therefore, piRNAs originating from them were suggested to play a role in translational regulation [91-93, 167]. Deletion of the MIWI2, one of the mouse Piwi homologues required for piRNA pathway, confirmed up-regulation of transposable

elements, which was attributed to loss of DNA methylation on their sequences [123, 168]. A similar phenotype was observed upon deletion of mouse MAEL [112]. Furthermore, in the absence of MAEL, the localization of MIWI2 and other piRNA-related proteins was disrupted, highlighting the requirement for this protein in context of piRNA pathway in cytoplasm [111]. Unlike the fruit fly, mouse MAEL localizes primarily to nuages in the cytoplasm, and its nuclear localization is obscure and most apparent in early embryonic stages [112]. Immunoprecipitation experiments of mouse MAEL show that it directly interacts with multiple piRNA pathway proteins, and complexes containing it are enriched with piRNA precursor and retrotransposon transcripts [154]. Despite of all these observations, the biochemical function of MAEL protein is still not very clear. In an effort to address this, two computational studies have identified the presence of an RNAse H-like fold in central Maelstrom-specific domain (MSD) [169, 170]. A presence of RNAse H-like domain, association with pre-piRNA transcripts, and lack of nuclear localization in the mouse would suggest that mouse MAEL may be involved in interactions with RNA instead of DNA. While this is consistent with the majority of fruit fly studies, some suggest that Mael may have additional nuclear functions [166, 171]. While the presence of amino-terminal high mobility group (HMG) –box domain required *in vivo* may suggest this, such functions were never experimentally supported [172]. Therefore, MAEL's prominent association with piRNA pathway and its domain composition paint a picture of where it is involved with cytoplasmic nucleic acids (RNA).

MAEL and Transposon Silencing

Transposable elements (TEs) are abundant in eukaryotic genomes [173]. First class, retrotransposons, propagate through an RNA intermediate in a copy-and-paste fashion utilizing chaperone, endonuclease and reverse transcriptase activities encoded within full-length elements [174]. Second class, DNA transposons utilize a cut-and-paste mechanism and their encoded transposase activity to move from one location in the genome to the next [175]. Both classes have been suggested to influence their hosts' genomes in positive and negative fashions [176, 177]. Retrotransposons, most notably a family of long interspersed nuclear element 1 (LINE1 or L1), take advantage of an epigenetic reprograming window during germline development and propagate [7]. The piRNA pathway discussed above functions to regulate transposon activity through reestablishment of the DNA methylation at their promoter regions [19, 26, 27]. Faithful silencing of transposons requires Piwi proteins to be loaded with piRNAs to aid in *de novo* DNA methylation of their sequences [7, 26, 123, 178]. piRNAs are generated from transposable elements and specific genomic regions, termed piRNA loci, both of which are transcribed by RNAP II like many other cellular RNAs [96, 178, 179]. Prominent localization of the piRNA proteins in cytoplasmic bodies suggests that these transcripts are processed there [111, 180]. However, how transposon and piRNA precursor transcripts are selected is not yet known. Loss of MAEL in mice leads to increased expression of transposons and defects in translational regulation that may be affected by unengaged piRNA machinery [112, 181]. Therefore, MAEL may facilitate steps required for engagement of piRNA proteins with RNA molecules to be processed into piRNAs.

Section summary

MAEL is the only protein in the piRNA pathway whose biochemical function is not yet described. Its domain composition suggests that it may be able to interact with nucleic acids, using either its HMG-box and/or MSD domain. However, the annotation of MAEL domains does not reveal the type of nucleic acids that it may bind. This is due to the prevalent role of HMG-box domains in binding to DNA and lack of evidence for RNA binding by MSD. Because of unsuccessful attempts to purify full-length protein, I have decided to focus my efforts on the HMG-box of MAEL, which is amenable to *in vitro* biochemical methods and for which solution structure (of human counterpart) has been determined [182]. HMG-box domains are capable of great sequence specificities and are capable of binding both types of nucleic acids (DNA and RNA). Ongoing characterization of the piRNA pathway is still devoid of important biochemical activities required in most presented models. The question interesting to me is, "what is the specific selection and processing of transposon as opposed to genic mRNA?" Processing of an appropriate transcript is not only important for a response to transposons through the piRNA pathway but also for unperturbed translation. Therefore, HMG-box of MAEL makes an attractive candidate for a domain that may preferentially recognize transposon RNAs and aid in their shuttling. This type of activity would constitute a novel role for HMG-box, antagonizing over-emphasized DNA-binding abilities of HMG proteins, and follow along the path of elucidating HMG-box functions in novel contexts.

High Mobility Group (HMG) of Proteins

The HMG proteins were fortuitously identified over 40 years ago by Johns group [183]. In an effort to obtain pure preparation of histone H1, the Johns group devised a way of removing proteins contaminating the histone fractions. The contaminants were characteristic in that they were extractable with 0.35 M salt, soluble in 2% TCA and traveled fast and often aberrantly on the acrylamide gels. It was this last observable characteristic that inspired their current name, "high mobility group" proteins [183]. HMGs are generally small proteins ranging in size from ~9 to ~30 kDa and have a high content of charged residues that contributes to their aberrant migration in gels [183]. After almost 30 years under scrutiny, a systematic way of classification and naming was proposed, subdividing them into three superfamilies: HMGA, HMGN, and HMGB [184]. I will briefly describe structural characteristics of HMGA and HMGNs and then focus my attention on the HMGB superfamily, which is most relevant to this thesis.

HMGA Superfamily

Proteins in the HMGA group are characteristic with an 'AT-hook' motif. This motif's name is based on its preferential binding to AT-rich regions in the minor groove of DNA. A single AT-hook is a short, palindromic motif enriched in charged residues (lysines -K, arginines -R) that exists as a random coil when unbound [185]. Canonical HMGA proteins usually contain three AT-hooks, however as many as 15 motifs have been identified in plants [186, 187]. Binding to nucleic acid stabilizes motifs` confirmation, revealing a concave structure that snugly fits into the minor groove of the DNA. The palindromic orientation of the residues in the groove introduces only minimal

bent to the DNA backbone. Core motif sequence is usually flanked with prolines (P), bending the peptide backbone away from the helix and giving the motif "hook"-like appearance [185]. The protein-DNA interactions are stabilized through H-bonds of polar side-chains with AT bases and more importantly, by their hydrophobic interactions with the phosphate backbone sugars [185, 188]. Members of this superfamily are HMGA1 and HMGA2. Both of them interact biochemically with a great number of proteins, making them relevant to diverse biological processes and diseases that have been reviewed elsewhere [189, 190].

HMGN Superfamily

Members of this group are HMGN proteins 1-5. All of them possess unstructured nucleosome binding domain (NBD) with a completely conserved sequence of 8 residues (RRSARLSA) required for binding to the nucleosome core *in vitro* and *in vivo* [191]. While NDB binds the nucleosome protein core, the acidic C-terminus gets modified and contacts histone tails, thus conferring regulatory and epigenetic functions onto HMGNs [192]. Their binding overlaps that of H1, suggesting that they may compete [193] and, in doing so, dynamically modulate the accessibility to DNA [194]. While HMGN proteins are ubiquitously expressed in adults, they play prominent roles during embryonic development, affecting transcription factor binding [195]. The diverse roles and functions of HMGN superfamily members have been extensively reviewed elsewhere [196-199].

HMGB Superfamily

The characteristic of the HMGB group of HMG proteins is a HMG-box fold. In this fold, three helices collapse into a L-shaped structure surrounding at the hydrophobic center with tryptophan residue (W). The tryptophan is highly conserved, and its mutation leads to loss of tertiary structure and concordant loss of specific binding; although a certain degree of nonspecific binding still remains [200, 201]. A single HMG-box has an unstructured N-terminus that continues into the first α−helix, which connects to the second α−helix via a short loop region. The amino-termini of these two helices commonly contain residues important to the domain's nucleic acid-binding abilities [202-204]. Helix one and two run in opposite directions to each other and form a short arm of the domain that is oriented roughly perpendicularly to the third, and commonly longest, α−helix. The domain fold concludes in an unstructured and often highly-charged C-terminus of variable length, depending on protein. A model of the candidate HMG-box domain is shown in Figure 3.

The unstructured termini perform important roles in substrate interactions, affecting them in either a positive or negative manner [202, 205-207]. They lack secondary structure in solution, however, when binding to a substrate, they get stabilized through the formation of multiple electrostatic contacts with substrate, also stabilizing the whole complex [208]. HMG-box localization and functions are modulated by numerous post-translational modifications, including acetylation, phosphorylation, or methylation, the majority of which are on residues in the unstructured termini and have been discussed elsewhere [209-213].

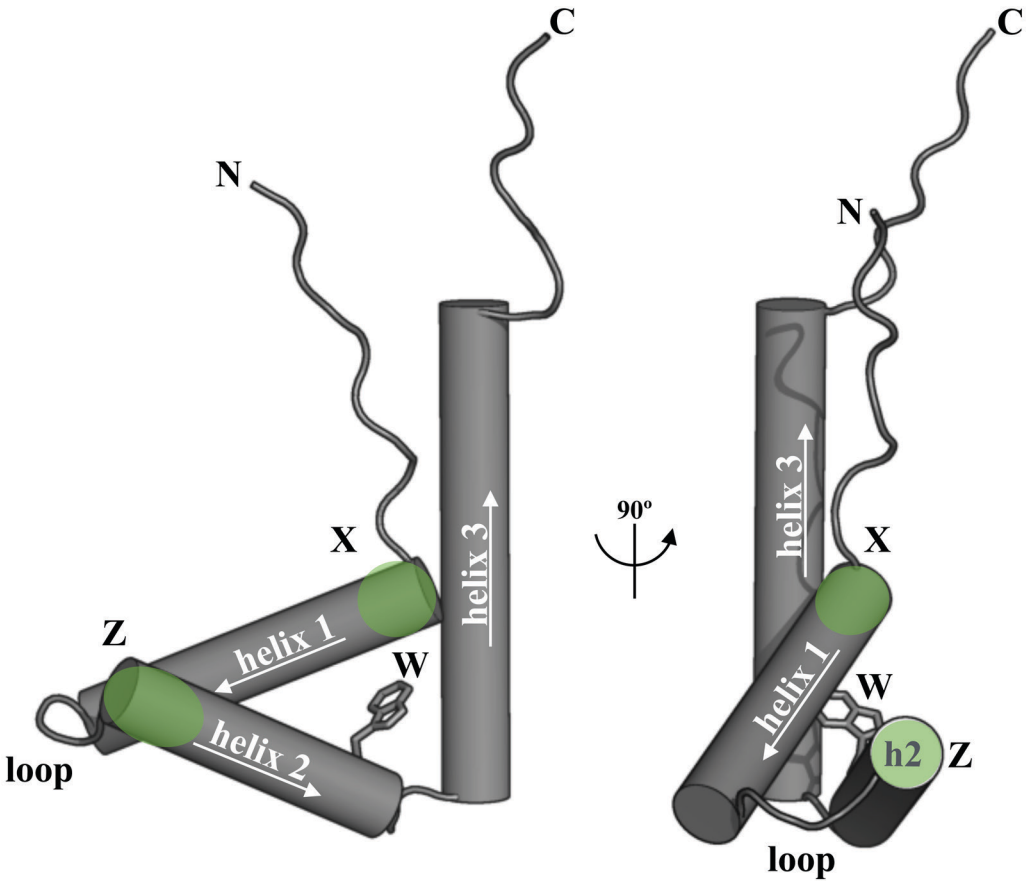


Figure 3: Tertiary structure of a HMG-box domain.

Three helices fold into a L-shaped structure encapsulating the tryptophan (W) in the hydrophobic center. The termini (N, C) are unstructured in the solution. Approximate locations of residues (X, Z) of primary importance to binding are highlighted in green. This figure is based on human SRY HMG-box (PDB ID: 1j46).

HMG-boxes and Double-Stranded (ds) DNA

HMG-box containing proteins can be divided into further categories based on the number of HMG-boxes present or their substrate preferences. The first group of HMG-box proteins contains two or more domains (for example: HMGB1-4, mtTF1, ABF2, *Drosophila* Dsp1, UBTF) and have little or no sequence preference [214]. Instead, these proteins bind well to perturbed and structured DNA such as modified by cis-platin (Pt) or found within DNA four-way junctions (4WJ) [215-218]. The second group contains single HMG-box domains and consists primarily of transcription factors (for example: SRY, LEF-1, SOX proteins) that bind to DNA in highly sequence-dependent manner [203, 219, 220].

Structural studies of both, sequence-specific (SS) and non-sequence specific (NSS) domains have led to the elucidation of their binding mechanisms and identification of residues that play key roles in complex formation. SRY HMG-box (SS) recognizes specific sequence in the target DNA and intercalates isoleucine (Figure 4A) residue between the bases of consecutive minor groove adenine bases, thus perturbing base tilt and roll and causing the backbone to bend. Perturbation of the DNA helical geometry leads to better accommodation of asparagine in the minor groove and formation of a stable complex (Figure 4A) [219]. On the other hand, box A of HMGB1 protein (HMGB1a, NSS) has alanine residues with side-chain too short to be able to intercalate between the bases; therefore, dsDNA backbone cannot be bent (Figure 4B). However, if dsDNA helix is already structurally perturbed, for example with a cis-Pt modification, the hydrophobic phenylalanine gets buried within the pocket, forming a stable complex (Figure 4B) [221].

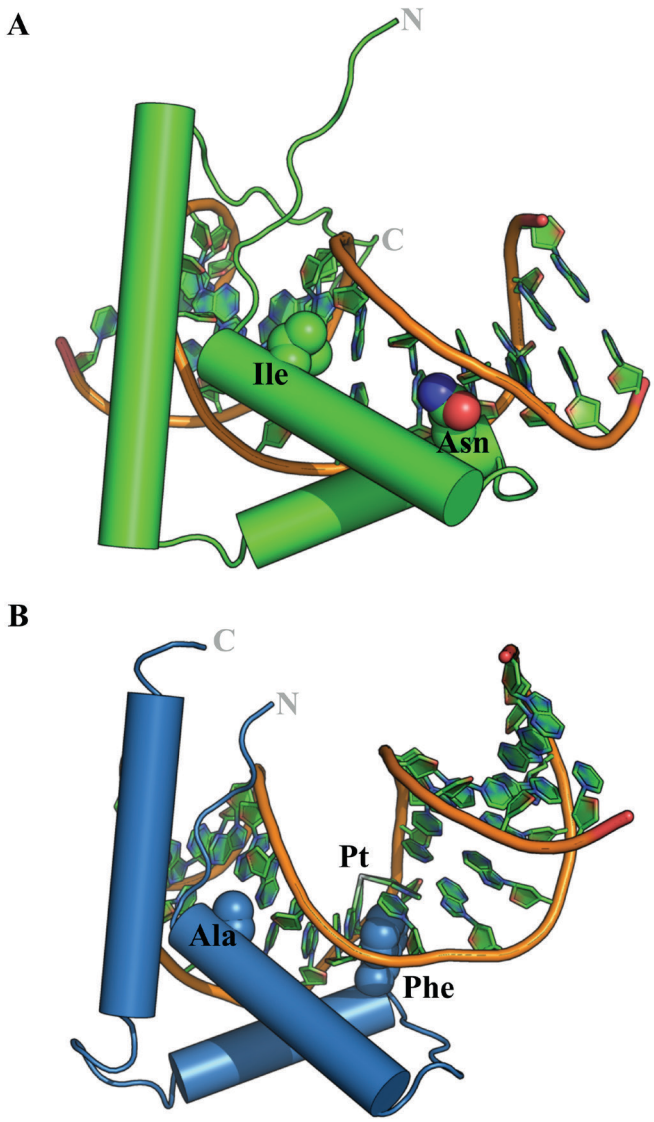


Figure 4: Interactions of two HMG-boxes with dsDNA

A) Sequence-specific interaction of SRY HMG-box with DNA containing its binding site. Isoleucine (Ile) and asparagine (Asn) are important for bending and binding to minor groove of dsDNA (PDBID: 1j46). B) Non-sequence-specific interaction of HMGB1 HMG-box A to dsDNA modified with platinum (Pt). Alanine (Ala) is too short to intercalate; instead, binding is dependent on hydrophobic burial of phenylalanine (Phe) in the pocket generated by cis-Pt modification (PDBID: 1ckt).

Multiple other residues along with termini are also involved in the stabilization of the final complexes for both of these proteins, however ones described above are of primary importance. In the case of HMGB1, the second HMG-box (box B) also contributes to its binding [222].

DNA Junctions as HMG-box Substrates

Aside from binding to regular and perturbed ds DNA, it appears as a natural ability of HMG-box domains to bind to DNA 4WJs [223, 224]. The central region of 4WJ is unlike B-type helical dsDNA because the strands are sharply bent and can cross each other. This geometry presents an attractive (electrostatically and hydrophobically) site for binding without requiring bending by the protein, especially when the junction is in the open conformation with ds portions extended in opposite directions [225, 226]. It has been shown that under low salt conditions, DNA 4WJ predominantly exist in the open conformations [227, 228]. The DNA 4WJs, also known as Holliday junctions (named after Robin Holliday who first described them), are important intermediates of homologous recombination during meiosis or DNA repair [229-231]. As such, many DNA 4WJ-interacting proteins were identified, some of which also interact with HMG-box proteins [232]. Directly or through protein-protein interactions, DNA 4WJs may be important biological substrates for HMG-box proteins.

HMG-box Binding to RNA

HMG-box domain-containing proteins have been shown to bind to ssRNA as well as complex RNA molecules [233-237]. Large RNAs contain various junctions, including

4WJs, which form in process of their tertiary structure formation [238, 239]. While synthetic DNA 4WJs exist primarily in open conformation, in solution the RNA counterparts are far more dynamic, mostly populating the closed structural ensemble [240]. An *in vitro* evidence shows that at least NSS HMG-box domains bind them well, suggesting these may be one of the structural features recognized in large RNA molecules [237]. Furthermore, the RNA-binding mode of NSS HMG-box proteins will be an intriguing subject to study in light of their involvement with recognition of immunogenic nucleic acids [241, 242]. Many viruses endemic to humans have RNA genomes and may be subject for binding HMG-box containing proteins [243].

HMG-box evolution

The diversity of context and functions within which HMG-box domains exist suggest that they have undergone extensive adaptive processes. The phylogenic analysis estimates that these domains arose over 1 billion years ago and have been rapidly evolving [244]. HMG-box domains are small, consisting of ~76 residues for the core fold, and are capable of sub-specialization, in some cases with acquisition of only few residue changes [220]. Their adaptation can occur independently of other domains in the same peptides thus providing great potential for sub-functionalization of proteins [244]. The majority of HMG-box domains described thus far associate with chromatin and serve as DNA-binding modules, and because of this, even ones in non-chromatin interacting proteins are construed to have evolved to accomplish DNA binding [166, 245].

Section summary

Proteins within the HMG superfamily accomplish a wide palette of functions, most of which entail chromatin and nucleic acids. Of the three subfamilies, roles of HMG-box group have been described in most detail, owing to the fact that their fold is most stable in amenable to *in vitro* techniques. The fast evolutionary adaptation of HMG-box domains to diverse cellular processes mandates rigorous validation to unambiguously describe their true functions. In this thesis, I will characterize the function of MAEL proteins' HMG-box, which appears to have diverged from canonical domains to contribute functions important to piRNA pathway.

CHAPTER II

MATERIALS AND METHODS

Phylogenetic Comparison of MAEL HMG-boxes

The Basic Logical Alignment Search Tool (*BLAST*) at National Center for Biotechnology Information (NCBI) was used to search for available Maelstrom nucleotide sequences. Initial query sequences, mouse Maelstrom (NM_175296.4) and fruit fly Maelstrom (NM_168958.3), were used to query nucleotide collection (nr/nt) and nucleotide sequences of Maelstrom from new species were extracted according to following criteria: containing ATG start codon, an unambiguous stop codon. Each one of those was then used as a query for a new search until no new Maelstrom sequences appeared in search results. As a results a total of 87 sequences were retrieved. The first 258 nucleotides corresponding to first 86 residues containing an HMG-box domain were then extracted and aligned using *ClustalW* algorithm in *MEGA6* software package [246]. The alignment was further and only sequences that passed following criteria were retained: presence of conserved tryptophan residue, presence of glycine at the end of first α-helix, presence of all three α-helices, no rearrangements or deletions larger than 10 consecutive residues. The labels of each sequence include corresponding accession number and common name of the organism. The alignment was adjusted to reflect secondary structural elements (α-helices and loops) using NMR structure of human MAEL HMG-box (PDB ID: 1j46) [182]. The resulting alignment consisted of 61 MAEL HMG-box sequences and was used for comparison of residue sequences and for phylogenetic analysis. The Phylogenetic tree was build using maximum likelihood method in *MEGA6* using following setting: test of phylogeny by Bootstrap method with 1000 replicates, amino acid substitution model using standard code table and Jones-Taylor-Thornton (JTT) model [247], rates among sites considered as gamma distributed (5 categories) with invariant sites (G + I), partial

deletion of gaps with site coverage cutoff of 80%, tree inference using Subtree-Pruning-Regrafting – Extensive (SPR level 5), and very strong branch swap filter. The highest log likelihood was -4531.75. The final gamma parameter (G) was 5.0811. There were 2.94 % invariable sites in total of 83 positions considered amongst 61 sequences. The tree branches and annotations were adjusted in the *MEGA6* built-in tree editor.

Comparison of MAEL and Canonical HMG-boxes

The full-length mouse Maelstrom sequence (434 residues) was submitted for tertiary structure prediction to the *Robetta* online server [248]. The .pdb files were retrieved and analyzed in *PyMOL*. The same process was followed for the *Drosophila melanogaster* Maelstrom HMG-box domain (residues 1-86). Nucleotide sequences of the candidate sequence-specific (SRY, SOX) and non-sequence-specific (HMGB, Dsp1) HMG domains were obtained from NCBI and 86 residue region encompassing HMG box was selected for the alignment. The sequence id indicates protein name + species + start residue + number of consecutive residues extracted. The accession numbers used were: Maelstroms –[*Mus musculus* (Mm) Maelstrom - NM_175296.4, *Drosophila melanogaster* (Dm) Maelstrom - NM_079493.4, *Homo sapiens* (Hs) Maelstrom - DQ076156.2]; sequence-specific HMG proteins [Hs SRY - X53772.1, Mm SRY - NM_011564.1, Mm Sox2 - NM_011443.3, Mm Sox6 - U32614.1, Mm Sox10 - AF047043.1, Mm Sox17 - NM_011441.4]; non-sequence-specific HMG proteins [Dm Dsp1 - U13881.1, Mm HMGB1 - NM_010439.3, Mm HMGB2 - NM_008252.3, Mm HMGB3 - NM_008253.3]. Codon alignment was performed using the *ClustalW* algorithm built-in *MEGA6* package, without changing pre-set parameters. The aligned

nucleotides were translated to protein using standard genetic code and the alignment of protein repeated using built in *ClustalW* algorithm. No changes to codon alignment occurred. This alignment was manually refined using experimentally determined structural information to account for the secondary characteristics such as helices and loops. Following PDB structures were used: SRY – 1j46 [219], HMGB1a -1ckt [221], MAEL HMG – 2cto [182]. The final alignment was exported and the residues colored according to Taylor color scheme to reflect biochemical characteristics of various residues [249]. The alignment was then used to generate maximum likelihood tree using *MEGA6* [246, 247] built-in algorithms with the following settings: 1000 Bootstrap replicates, Jones-Taylor-Thornton model of amino acid substitutions, uniform site-rates, complete deletion of gaps and missing data, Subtree-Pruning-Regrafting – Extensive at level 5, very strong branch swap filter. The generated tree was visually adjusted in built-in tree editor and is presented in Figure 10A. The log likelihood of this tree is -1943.6 and each branch is annotated with the bootstrap values representing the percentage of trees where the associated sequence clustered together. The tree branch scale represents number of substitution per site based on the considered 73 completely conserved positions amongst 17 compared sequences.

Cloning and Mutagenesis

Mouse Maelstrom cDNA was previously generated in the lab [112]. HMGB1a cDNA was obtained from Open Biosystems (clone 30849071). SRY HMG was PCR amplified from mouse testis 129S4 cDNA. *Drosophila melanogaster* maelstrom was amplified from *Drosophila* testis cDNA (gift of the X. Chen lab, Johns Hopkins University). All

HMG domains (nucleotides 2-258 coding for first 85 residues) were amplified with Phusion polymerase (NEB) using primers listed in Table 2. PCR products were sub-cloned into pGex6P2 expression vector (GE) between BamH1 and Not1 restriction sites. Selected residues were mutated using round-the-horn site-directed mutagenesis [250] and confirmed by sequencing.

Protein Expression and Purification

Domains sub-cloned into pGex6P2 vector were transfected into BL21-DE3* cells (Life Technologies). Single clones were expanded overnight and inoculated into a large volume of terrific broth (TB) media supplemented with appropriate antibiotics. When the OD_{600} reached 0.6-0.8, the culture was moved to 18°C incubator and protein expression was induced with 250 mM IPTG for 12-16 hours. The cells were collected following day by centrifugation, washed once with 1xPBS and resuspended in lysis buffer (1x PBS, 5% glycerol, 1mM PMSF, 1mM TCEP, 1mM $MgCl_2$, protease inhibitors (Pierce)). The cell suspension was supplemented with Lysozyme (Sigma) to the final concentration of 1-2 µg/ml and incubated on ice for 30 minutes with occasional mixing. Lysed material was then sonicated (4 repeats, 20 second sonication, 50% duty, Misonix 3000) with one-minute incubation on ice between repeats. The sonicated mixture was spun (4°C, SS-34 rotor, 30 minutes, 18000 rpm) and the supernatant carefully moved to syringe and filtered with Millex HV filters (Millipore) to remove contaminants. The GST-fusion protein was purified by gravity on the glutathione agarose resin (Sigma) at 4°C unless otherwise noted. The filtered lysate was bound to the glutathione resin, washed with 5 column volumes of low salt buffer (LSB: 1xPBS, 5% glycerol), 5 column volumes of high salt

buffer (HSB: 1xPBS, 5% glycerol, 1M NaCl) and again with 2 column volumes of LSB. The protein was eluted with 10 mM reduced glutathione in LSB (pH ~8.5). To remove the GST tag, the eluate was supplemented with 1mM EDTA, 1mM TCEP, PreScission protease (GE) and incubated at 4°C for 12-16 hours. Phospho-cellulose (PC) columns (2.5 ml) were prepared from dry PC resin (Whatman P-11) following procedures provided by Lorsch Lab (NIGMS). Briefly, 0.8g of the resin was stirred into 125 ml of 0.5N NaOH for 5 minutes. After that the resin was washed with water until pH < 11 at which point 125 ml of 0.5N HCl was added and the solution was stirred for 5 minutes again. The mixture was then washed with water until pH > 4, at which point the resin was poured into disposable columns and equilibrated with desired buffer until $pH_{IN} = pH_{OUT}$. The PreScission digested glutathione column eluate was diluted with B0 buffer (B0: 20mM Hepes pH 7.4, 10% glycerol, 2 mM DTT, 0.1 mM EDTA) to lower the salt below 75 mM. The diluted digest was then loaded onto 2.5 ml PC columns and allowed to bind by gravity. The column was washed with 5 column volumes of B100 buffer (B0 + 100mM NaCl). The protein was eluted with B500 buffer (B0 + 0.5M NaCl). The fractions with A_{280} > 0.05 were pooled and their buffer exchanged (PD-10 desalting columns) to storage buffer (SB: 1x PBS, 5% glycerol, 1 mM TCEP). To remove residual PreScission protease and un-cleaved protein, the eluate was passed over 0.5 ml glutathione/0.5 ml Ni-NTA column. The final eluate was concentrated using Vivaspin 6 centrifugal concentrators (Satorius) with 3 MWCO to desired protein concentration (>1 mg/ml). The concentration was measured at A_{280} in 6M Gnd-HCl, 20 mM Sodium Phosphate pH7.5 using calculated extinction coefficient and molecular weight (http://www.expasy.org) on NanoDrop 2000c. The final protein was aliquoted, flash

frozen and stored at -80°C until use.

Circular Dichroism (CD) Spectroscopy

CD measurements were collected on Aviv 420 instrument (Aviv Biomedical). Far-UV spectra were collected in 0.1 cm cuvette at 25°C. All proteins were 94-residues long at 0.1mg/ml concentration. The Samples were in 1x PBS, 5% glycerol at room temperature. The data were processed in Numbers (*iWork*, Apple Inc.) as previously described [251].

Simple Substrate Preparation

The DNA oligonucleotides for each substrate were purchased desalted without purification (Operon) (Table 3). The RNA oligonucleotides were ordered desalted with HPLC purification (Sigma Proligo) or prepared by *in vitro* transcription of PCR products with T7 promoter using HiScribe T7 high yield RNA synthesis kit (NEB) following manufacturers` protocol (Tables 4, 5, 6). Briefly, the synthesized RNA was purified with acid phenol:chloroform and precipitated with isopropanol. The precipitate was diluted in 1x CutSmartTM buffer (NEB), supplemented with RNAseIN inhibitors (Ambion) and de-phosphorylated with alkaline phosphatase (NEB). The precipitation was repeated. The precipitate was then purified on TBE-UREA polyacrylamide gels (Live Technologies) and the RNA was purified following crush-and-soak method [252]. Briefly, the appropriate size band was excised from the gel, crushed in the presence of PAGE elution buffer (0.3M NaOAc, 10 mM Tris 8.0, 1 mM EDTA pH 8.0) and frozen at -80°C for 30 minutes. The RNA was eluted by shaking the mixture overnight at 37°C and precipitated with isopropanol. The molar concentration was calculated based on the A_{280} readings in

10 mM Tris pH 8.0. The RNA was stored at -80°C until use.

Complex Substrate Preparation

The oligonucleotides were designed with sufficient overlap and homology to specifically anneal. To create a four-way junction, 10 µl (100 µM) of each oligonucleotide was mixed in the annealing buffer (1x: 70 mM Tris pH 7.5, 10 mM $MgCl_2$, 100 mM NaCl) to a final volume of 200 µl. The mixture was incubated in 95°C water bath for 5 minutes and allowed to slowly cool to room temperature. The annealed substrate was precipitated with EtOH (DNA) or isopropanol (RNA). Approximately 10 µg of annealed substrate was diluted in the binding reaction without protein (1x: 10 mM Potassium Phosphate pH 7.5, 50 mM KCl, 5% glycerol, 1 mM TCEP, 2.5 mM $MgCl_2$), loaded into single lane of 12% Native page gel and ran at 105V for 1-2 hours. Lower concentration of the acrylamide was used for the RNA substrates > 75 bases. The bands were visualized using short wavelength UV shadowing and appropriate bands were excised and purified following crush-and-soak method described earlier. The molar concentration of each substrate was calculated using molecular weight and A_{280} readings. The oligonucleotides were aliquoted at desired concentration and stored at -80°C until used. All double stranded (RNA, DNA) substrates were annealed and purified in the same fashion. The RNA oligonucleotides for hairpin substrates were ordered HPLC-purified (Sigma Proligo).

RNA Substrate Structural Considerations

To simplify interpretations, all the substrates made from ssRNA were designed with potential secondary structural characteristics in mind. The sequences were submitted to

the *Mfold* server [253], using standard settings to identify thermodynamically favorable confirmations. All structures with negative free energy (ΔG) were considered as likely within the ensemble of tested RNA. The structures with +ΔG were considered as unlikely. This is based on the fact that base pairing provides -ΔG to RNA molecule allowing for spontaneous folding and secondary structure formation [254]. Therefore, in Mfold analysis, sequences that produce structures with only +ΔG are considered single-stranded, whereas an ideal hairpin sequence would produce only single structure with large –ΔG. The Table 4 contains the free energies and structures of the tested substrates identified by *Mfold*.

Gel Shift Assays

The substrates were diluted to desired concentrations and end-labeled with γ-P32 using PNK (NEB). To account for number of ends, 5 μM of the hot ATP were used per 1 μM of DNA four-way junction. Unincorporated label was removed on P30 columns (Bio-Rad). To control for the loss of the shorter substrates on the P30 column, multiple substrates were labeled at the same time and their concentrations were normalized to the DNA 4WJ (largest substrate) using relative incorporated scintillation counts. Such prepared substrates were stored at 4°C until use, unless folding was required. To fold, the RNA substrates were supplemented with salts (50 mM NaCl, 2.5 mM $MgCl_2$) and heated to either 55°C (< 50 bases) or 95°C (>50 bases) for 3 minutes and allowed to slowly cool to RT. The folded RNA was stored at 4°C until use. The protein was thawed on ice and then serially diluted to desired concentrations in water. The binding reaction was assembled by mixing the protein in binding reaction consisting of (1x) 10 mM Potassium Phosphate

7.5, 50 mM KCl, 5% glycerol, 1 mM TCEP, 0.1 mg/ml BSA, 2.5 mM $MgCl_2$. The labeled substrate was added last to ~1nM concentration in 10 µl final volume. The reaction was then incubated at room temperature for 30-60 minutes to equilibrate. The 12% native polyacrylamide (29:1), 1 mm thick TBE (1x) mini-gels gels were pre-run wit 0.5x TBE running buffer for 30 minutes at 105V in ice water-bath. The wells were briefly rinsed, and 5 µl of the binding reaction was then carefully loaded onto running gels. The gels were run at constant 105V for long enough (1-4 hours) to achieve sufficient complex separation. At the end of the run, gels were extracted, rinsed, dried onto 3 mm Whatman paper at 80°C for 90 minutes, and exposed to storage phosphor screen for 12-24 hours. The image was acquired using Storm 860 molecular imager (Molecular Dynamics) with 100-micron resolution. The large RNA substrates were treated the same way but the complexes were resolved on large 6% native gels. In the competition experiments, the binding reactions were setup in the same manner as above but with protein concentration held constant and sufficient to achieve between 60-90% binding. Serially diluted unlabeled (cold) substrate was added up to 1 uM final concentration prior to addition of the radioactively labeled (hot) substrate.

Data Analysis

The images obtained from the Storm 860 were analyzed in *FIJI* (GPL). The region of the gel was extracted, the pixels inverted onto black background, and background subtracted uniformly amongst all images. For each lane the region free and the region bound were selected using gel analysis feature and the area under the curve quantified using wand tool. Multiple complexes were all included in the region bound. The fraction bound was

calculated using equation (1) and data plotted as the fraction bound versus protein concentration using *Prism6* software (*GraphPad*). To calculate dissociation constant (K_D), data was fit to modified Hill equation (2). The cold competition data was plotted as the fraction of bound hot substrate versus the concentration of the cold substrate using equation (3) and the dissociation constant of competitor (K_C) was calculated with equation (4). All parameters in equations (2,3,4) were described previously [255].

$$Fb = bound/total \quad (1)$$

$$f = b + [(m - b) / (1 + (K_D / [P_t])^n] \quad (2)$$

$$f = b + [(m - b) / (1 + (IC_{50} / [C])^n] \quad (3)$$

$$K_C = (2K_D IC_{50}) / (2P - R - 2K_D) \quad (4)$$

Large RNA Structure Determination

Previously described MAEL RIP-Seq data sets mapped to mm9 assembly of the mouse genome, shown to be enriched in transposon RNA, were used for the identification of over-represented regions [154]. The sets corresponding to control Igg, MAEL_A RIP and MAEL_B were analyzed with *macs* software (version 1.4.2) [256] with the standard settings to identify regions enriched in replicates A and B over Igg. The identified regions between the two replicates were pooled and intersected using *bedtools* (v2.20.1) [257] to identify only the common regions. All intervals were then annotated using annotatePeaks program from *HOMER* suite [258]. The regions annotated as LINE1 elements were extracted and their coordinates examined in *IGV* [259, 260], considering only the regions within annotated LINE1 elements. Multiple coordinates corresponding

to regions with a peak appearance at least 250 nucleotides-wide were selected, and their nucleotide sequences extracted from the UCSC genome browser. These were then aligned using *ClustalW* (EMBL-EBI) and the alignment manually curated until the region of high sequence conservation was identified. The final alignment had 5 regions corresponding to LINE1 elements of Md_F2 family that were located on different chromosomes. Coverage across each identified region was calculated using its coordinates and the bedtools *multicov* program [257]. The results were plotted in Numbers (*iWork*, Apple Inc.). This alignment was used for determination of the secondary structure according to previously described methodology combining RNAalifold and *Mfold* [253, 261, 262]. The covarying nucleotides used to constrain *Mfold* are provided in Table 7. The region with lowest dG (chr10) was tested in gel shift assays.

CHAPTER III

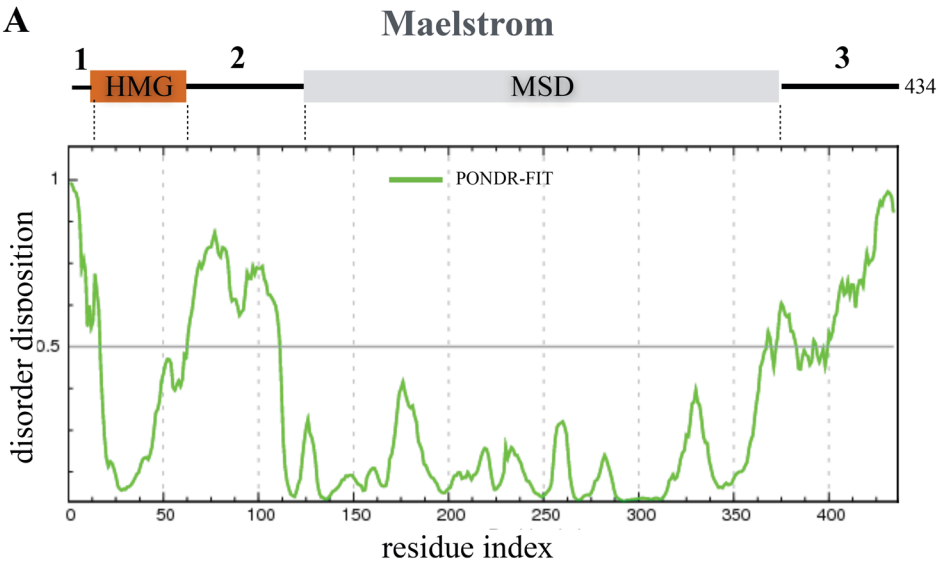
RESULTS

Insights From MAEL Amino Acid Sequence

Murine MAEL protein is 434 amino acids long and is annotated as having two domains: an HMG-box (*UniProt*, residues 1-76) and a novel maelstrom-specific domain (MSD, 100-434) [169]. However, regions that may be important for functions and regulation of the full-length protein and its domains have not been described. I have used computational prediction tools to identify regions with high probability for disorder (*PONDR-FIT* [263]), and then I analyzed their compositions with a focus on residues that can be phosphorylated (*NETPHOS2.0* [264]). The prediction of intrinsic disorder identified three regions (1, 2, 3) with probability of being disordered being greater than 50% (Figure 5A).

The first region (1) in the N-terminus is short and consists of 9 residues (Figure 3B). The amino-termini of HMG-box domains are commonly disordered and participate in binding [205]. This region contains positive (arginine, R) and polar (asparagine, N) residues that can form hydrogen (H) bonds with nucleic acid bases or the nucleic acid backbone [201, 221, 265]. After the HMG-box binds to nucleic acid substrate, the unstructured N-terminus becomes well-defined through interactions with nearby substrate regions [219, 221]. On the other hand, unstructured N-termini can be post-translationally modified (PTM), thereby modulating binding strength or protein localization [199]. An *in silico* prediction scan for phosphorylation potential in the N-terminus of MAEL HMG-box identified single serine (S) with high probability (99%) to be modified (Figure 3B). The unstructured nature (Figure 3A) and the residue content (Figure 3B) indicate that region 1 may influence binding of the whole domain either directly or potentially through PTMs.

A



B

	residue sequence	size	% P/G	% S/T/Y
region 1	MPNRRASRN	9	0	1.3
region 2	PSEKQKLVSTPLRP GMLVPKPSISPPDM SN	31	**25.8**	19.4
region 3	GITRLLESISNSSNN IHRFSSCETSLSPYT PQKDGYKPFSSFS	43	11.6	**37.2**

Figure 5: Disordered regions of MAEL

A) Predicted intrinsic disorder (PONDR-FIT server) across 434 residues of MAEL protein reveals three (1, 2, 3) regions with >50% disorder disposition. B) The residue composition within predicted disordered regions and fraction of residues associated with unstructured regions (P/G – counted manually) or are post-translationally modified (S/T/Y - NETPHOS 2.0 server). Phosphorylation probabilities >50%: region 1 - S7 = 99%; region 2 - S93 = 64%, S95 =99%; region 3 - S399 = 94%, S411 = 99%, S412 = 97%, Y427 = 88%. Colors correspondence: red >90%, purple >80%, blue>60%.

The second unstructured region (2) is 31 residues long and connects HMG-box domain with the MSD domain (Figure 3b). This region contains a high proportion (~25%) of residues with low propensity for secondary structure formation (prolines-P and glycine-G) and residues that are most commonly subjects of PTMs (~19% serine-S and threonine-T) [266-268]. This region is C-terminal to HMG-box domain in MAEL protein and is unstructured (Figure 5A). Like C-termini of other HMG-boxes, it may affect nucleic acid binding abilities [205]. Unstructured regions of various proteins contribute to their structures and functions by enabling energetically inexpensive motion of domains respective to each other [269, 270]. Such regions are also likely to play important roles in protein-nucleic acid and protein-protein interactions through either contributions to the binding affinity or by modulating the binding with acquisition of PTMs such as phosphorylation [271, 272]. The fruit fly Mael has been shown to be phosphorylated on nearby S138 with functional consequences [273], however mass spectrometric (MS) analysis of MAEL carried out by others in our lab did not focus on identification of PTMs; therefore, I cannot exclude a possibility that residues within unstructured linker carry some PTMs [154]. Using computational prediction, I detected two residues with >50% chance of being phosphorylated within this unstructured region (Figure 3B).

C-termini of many proteins are often unstructured, commonly contain multiple PTMs and can contribute to protein stability and regulation [274-277]. The C-terminus of MAEL (region 3) is 43-residues long and also predicted to be disordered (Figure 5A). It contains multiple residues that can be phosphorylated, most of all in three discussed regions (37.5% of all residues, Figure 3B) [264].

Section summary

The peptide sequence offers clues about MAEL characteristics that may be required for its function or regulation. The three described regions contain residues with potential to affect structure, function, or regulation of this protein. The N-terminal region (1) may play an important role in the context of the HMG-box, while the other two regions have either a high fraction of structure disrupting or PTM-modifiable residues in unstructured linker (region 2) and C-terminus (region 3), respectively, that may affect the whole protein. My experiences with purification of MAEL support this prediction. Constructs of full-length MAEL (GST and 6His N-terminal fusions) were insoluble in bacteria and insect cells, but after removal of approximately 30 residues from the C-terminus, the protein became more soluble. However, the small amounts of protein obtained aggregated as soon as the N-terminal affinity tags were cleaved, preventing further purification and analysis. The solubility of the two domains expressed separately further increased, and only minor aggregation occurred. Therefore, the unstructured C-terminal (3) and internal linker (2) regions likely contribute to full-length protein stability and solubility. It may be that outside of its native cellular environment these regions lack chaperones, PTMs, or binding partners that could stabilize them in the di-domain protein context, which is supported by high content of structure disrupting and modifiable residues.

The most intriguing of the identified regions is the unstructured linker (region 2) connecting HMG-box domain to MSD domain (Figure 5A, B). It is 31 residues long, and theoretically this peptide could extend up to 109 Angstroms (10.9 nm) based on published α-carbon distance in peptide chain (~3.5 A) [278, 279]. In context of RNA,

this corresponds to a distance spanning 36 base-pair long dsRNA [280]. Having a flexible arm capable of such extension can provide a great degree of flexibility for the di-domain MAEL protein and may be of functional significance. I believe that future studies of this and other MAEL-disordered regions may provide great insight into its regulation and function.

Insights From Prediction of MAEL Tertiary Structure

The full-length MAEL protein insolubility in bacterial host precludes *in vitro* study of this protein. Since the tertiary structure has not been solved thus far, I have turned to computational methods in hope to gain further insight into the function of MAEL through its tertiary structure. Computational algorithms for determination of protein structure are under continuous development and can reasonably approximate tertiary structures [281]. Therefore, I have used the *Robetta* server, one of the most accurate algorithms, to predict MAEL tertiary structure [248]. Robetta uses homology of previously determined structures in a protein data bank (PDB) in combination with *de novo* methods to determine tertiary structure of peptide sequence.

When *in silico* folding was completed, I recovered five lowest-energy structures and further examined one with the highest proportion of secondary structural features (Figure 6A). In accordance to Maelstrom gene annotation (UniProt) and previous analysis, the structure was annotated with an HMG-box in its N-terminus and MSD domain c-terminal portion [169, 170]. Structure of MAEL HMG-box (residues 1-92 in Figure 4A) is based on a previously solved H-NMR structure of human MAEL HMG-box protein (PDB ID: 2cto) [182]. The MSD domain has been previously described to possess an RNAse H-like fold [169, 170]. *Robetta* has used a structure of an exonuclease (PDB ID: 1zbh –chain A) as a parent molecule to model residues 92-277 [282]. Residues 278-434 were modeled with *de novo* algorithms (Figure 6A). The predicted structure shows that the HMG-box domain is on the surface and not encapsulated by the rest of the molecule. The MSD domain is predicted to have two lobes; structure of the first one is based on the above-mentioned endonuclease domain, and the second one is determined

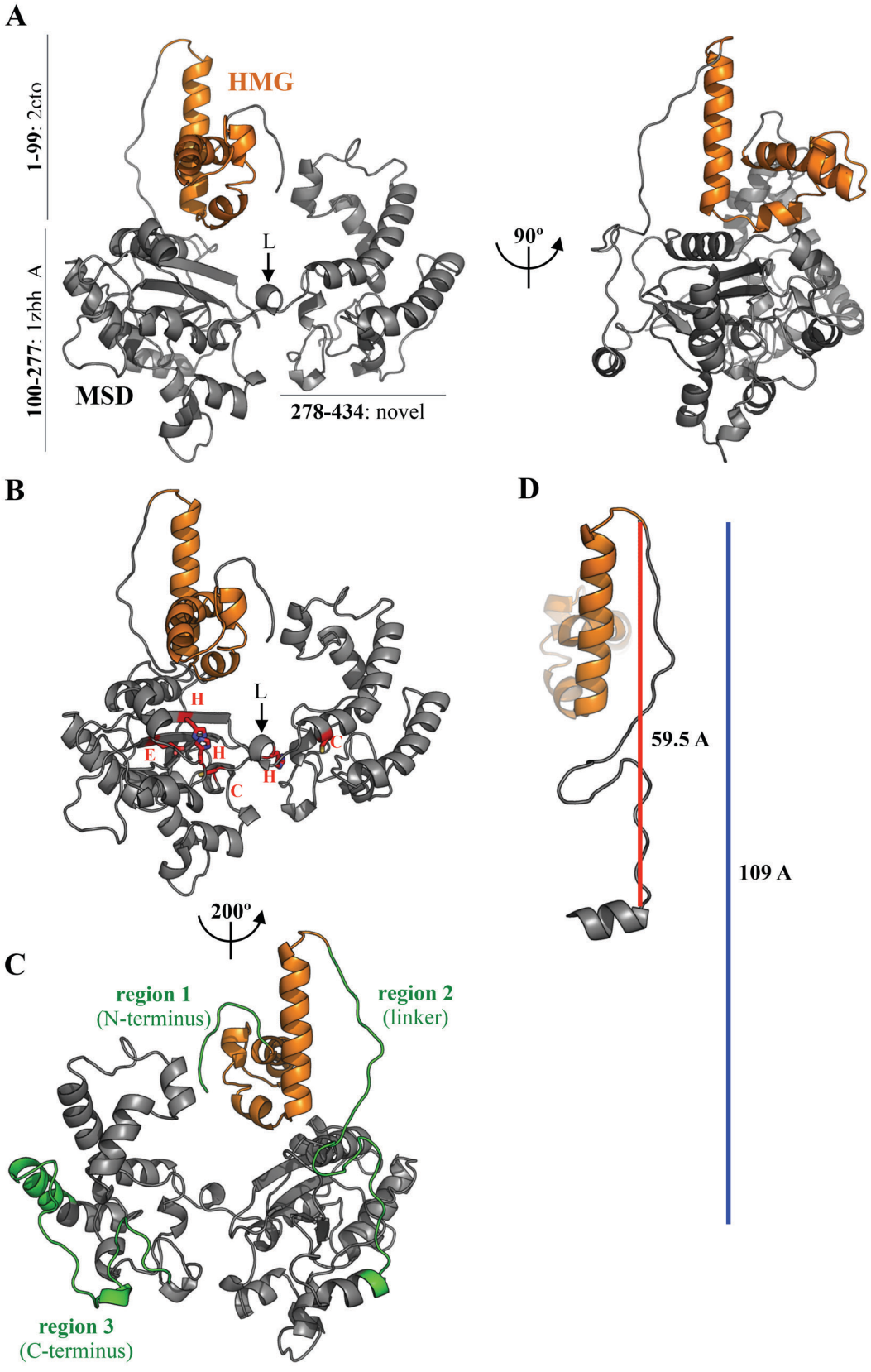
de novo. The two lobes are connected with a short linker (L in Figure 6A) of 9 residues (CKWHEENDI), the majority of which are hydrophilic, implying they may not need to be buried but instead may allow for some degree of movement of the two lobes respective to each other. This linker is the only peptide segment connecting the two lobes of MSD and is near the region that contains a highly-conserved set of residues (EHHCHC in Figure 6B) potentially related to a putative active site [169]. Multiple nucleic acid interacting proteins have channels near binding sites or leading up to catalytic sites [283-285]. In accordance with my sequence analysis, the predicted MAEL structure also shows the presence of three unstructured regions (Figure 6C).

The first, short amino-terminal region (1) is fully unstructured in the prediction (Figure 4A) as well as in the solution structure of the human MAEL HMG-box [182]. It contains multiple charged residues that are not buried within the structure (Figure 5B) and may be important for the HMG-box function [205, 219].

The MAEL domains are connected by a 31-residue linker, which is disordered in the predicted MAEL structure (2) (Figure 6C), also confirming my previous predictions (Figure 5A). This peptide sequence could theoretically extend to a distance of approximately 109 angstroms based on the maximal observed distance of the alpha-carbons in the peptide chains [278, 279]. In this predicted structure, the unstructured linker spans 59.5 angstrom and lacks secondary structural elements (Figure 6D). Because crystal structures and structural predictions capture the most energetically and structurally stable confirmations [286], in the native environment, this region could extend to the calculated limit and shift the position of the HMG-box relative to the rest of the molecule.

Figure 6: Tertiary structure of MAEL

A) Predicted tertiary structure of full-length MAEL (Robetta). The PDB IDs of parent molecules for peptide regions of the HMG-box (orange) and the MSD (grey) are shown. Highlighted with an arrow is the linker connecting two lobes (L: CKWHEENDI) within MSD. B) Location of conserved residues in MAEL according to previous predictions [169]. Residues are shown as sticks and are designated with one letter code where: E – glutamic acid, H – histidine, C – cysteine. C) Unstructured regions of MAEL identified by previous sequence analysis (Figure 3A) are present in predicted tertiary structure. D) The observed length of the HMG-box and MSD unstructured linker region (red) and to scale calculated maximal possible extension of the 31 residue long peptide (blue).


A
HMG
1-99: 2cto
100-277: 1zbh A
L
MSD
278-434: novel
90°
B
H
L
E
H
C
H
C
200°
C
region 1
(N-terminus)
region 2
(linker)
region 3
(C-terminus)
D
59.5 A
109 A

Lastly, the C-terminal region of MAEL is predicted to be largely unstructured based on its sequence (3) (Figure 5A). This region is also unstructured in the predicted tertiary structure, however not completely, with portions forming α−helix (Figure 6C). The structure of residues 278-434 was predicted without a parent molecule using *de novo* methods, indicating that this portion of MAEL has no structural homologues. Because the sequence and structural prediction methods can be wrong, additional biochemical methods are required to identify the true structure of this region.

Section summary

Overall, the *in silico* structural analysis of full-length MAEL agrees with our sequence considerations. The HMG-box structure is based on human MAEL HMG-box solution structure and therefore best approximates its real structure. However, MSD is partly based on the solved structure of an unrelated exonuclease and is partly estimated. Both sequence and structure analysis indicate that the N-terminus (region 1) and the linker connecting the two domains (region 2) are both unstructured, however, unlike for sequence analysis, the c-terminus is partly folded using structural prediction. Because both of these analyses rely on prediction methods lacking perfect accuracy, they require experimental validation and are to be considered with caution. In the following sections, I will focus solely on the MAEL HMG-box domain whose structure has the strongest experimental support.

MAEL HMG-box Domain is Ancient and Highly Conserved

MAEL is the only HMG-box domain containing protein in the piRNA pathway, and while it plays important biological roles in this context as determined by mouse genetics [111, 112, 154], biochemical functions of its domains have not been described thus far. As discussed earlier, HMG-box domains have a typical L-shaped fold (Figure 3), and in most HMG-box domains containing proteins, they constitute an efficient DNA-binding module [219, 220]. However, the association of MAEL HMG-box with MSD, a domain with predicted RNAse H-like fold [169, 170], as well as MAEL involvement in the piRNA pathway that uses small RNAs to silence retrotransposons, all suggest that its function may have diverged from those of canonical HMG-boxes towards functioning in context of RNA. Amino acid sequence and corresponding protein structure can provide valuable insights into protein's biochemical function. Therefore, in the following sections, I will discuss sequence, structural, and evolutionary considerations of the MAEL HMG-box domain and contrast them with canonical HMG-box domains.

Human and Mouse HMG-boxes

The solution of H-NMR structure of human MAEL HMG-box was solved by a consortium at the RIKEN Institute but without any additional characterization [182]. In the previous section, I discussed the computationally predicted structure of full-length mouse MAEL (Figure 6) from which I extracted a portion corresponding to the HMG-box domain. To establish a degree of similarity between the mouse and human domains, I compared their sequences and structures. Based on pair-wise comparison of their protein sequences, the mouse and human proteins are highly conserved with 91.9% identity.

Within the first 86 residues only 7 amino acids were different, and the majority of differences conserve the chemical properties of their side chains (Figure 7A). A high degree of sequence conservation implies a high degree of structural conservation. Accordingly, the structural alignment of the human domain, which has been determined by NMR, with the predicted mouse domain failed to show any differences in their secondary structural characteristics (Figure 7B). The residues that are different between the two domains do not cause any structural perturbation, as would be expected based on conserved sizes and types of their side-chains (Figure 7B). Both sequence and structural similarities suggest that the study of mouse domain function would directly apply to its human counterpart.

MAEL HMG-box Homologues

To gain an insight into the evolution and conservation of MAEL HMG-box, I retrieved available MAEL sequences from National Center for Biotechnology Information (NCBI). Many sequences in NCBI are a result of new genome sequencing with subsequent open reading frames (ORFs) identification by automated algorithms. Therefore, I have considered only sequences whose ORFs included a start codon, appeared to have a HMG-box, and had an unambiguous stop codon. The first 86 residues (containing the HMG-box) of each sequence were selected and used for multiple sequence alignment, and sequences that did not pass the following criteria were eliminated: the presence of all three α-helices and a core tryptophan (W) required for proper folding, and overall conservation of topology typical of HMG-box domains [201, 202, 208, 214, 222, 245].

A

```
                          1 [        .         .         .         .         : 50
1 MmHMG_1-86      100.0%    MPNRRASRNAYYFFVQEKIPELRRRGLPVARVADAIPYCSADWALLREDE
2 HsHMG_1-86       91.9%    MPNRKASRNAYYFFVQEKIPELRRRGLPVARVADAIPYCSSDWALLREEE
  consensus/100%             MPNR+ASRNAYYFFVQEKIPELRRRGLPVARVADAIPYCSuDWALLRE-E

                         51         .         .         .     ] 86
1 MmHMG_1-86      100.0%    KEKYSEMAREWRAAQGKDSGPSEKQKLVSTPLRRPG
2 HsHMG_1-86       91.9%    KEKYAEMAREWRAAQGKDPGPSEKQKPVFTPLRRPG
  consensus/100%             KEKYuEMAREWRAAQGKDsGPSEKQK.V.TPLRRPG
```

B

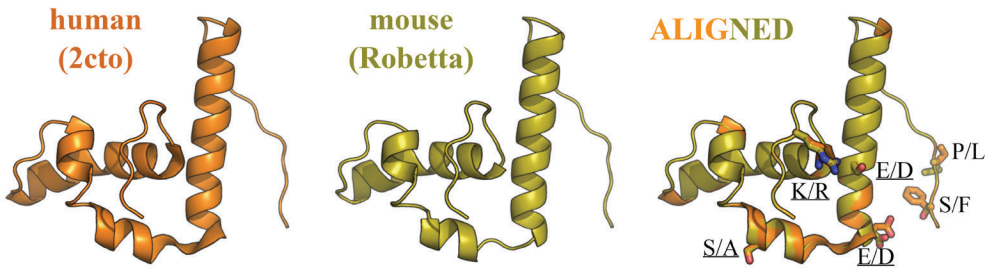


Figure 7: Comparison of human and mouse MAEL HMG-box

A) Sequence alignment showing high residue identity (91.9%) between the first 86 residues of human and mouse domain. The residues that are different have conserved side-chain properties. Consensus: + positive, **-** negative, **u** tiny, **s** small, **.** not matching. B) Structural comparison of human MAEL NMR structure (PDB ID: 2cto, orange) and mouse structure predicted by *Robetta* (teal). The secondary structural elements are unchanged by residues that are distinct between the two domains.

The final set of sequences was aligned (*ClustalW* in *MEGA6*) and adjusted for structural elements using solution structure of human MAEL HMG-box [182]. The final alignment contains the HMG-box of 61 MAEL proteins from a variety of species (Figure 8). At first glance, most striking is the very high residue conservation of the mammalian MAEL HMG-boxes. Of the first 86 residues, there are only minimal differences, most of which are concentrated toward the c-terminus. On the other hand, the most divergent sequences belong to HMG-boxes found in insects. These appear to have a shortened N-terminal region as well as several charged residues within loops and helices that are different from other animal classes (Figure 8). Other represented classes (birds and reptiles) appear to have sequence characteristics more similar to mammals than to insects, i.e., the length of the amino-terminus and charged residue distribution in birds and reptiles is very similar to that seen in mammals (Figure 8).

To better understand the relationships between MAEL HMG-box sequences, I performed phylogenetic analysis using the *MEGA6* package [246]. The resulting phylogenetic tree describes the relatedness of the HMG-box domains based on their residue composition and is calculated using the maximum likelihood (ML) method (Figure 9). The mammalian domains (class mammalia) all group together, and their closest relatives are reptilian and bird (class aves) domains. Insect domains (class insecta) are most distantly related to the first two groups, and divide into different branches corresponding to orders of diptera (drosophila, mosquito) and heminoptera (bees, wasp, aphid) (Figure 9). Few species from multiple insect groups (southern house mosquito, pea aphid, monarch butterfly) do not cluster with other members of their order but are still found within the appropriate insect class (Figure 9).

Figure 8: Sequence comparison of multiple MAEL HMG-boxes

A multiple protein sequence alignment of the HMG-box domains from 61 selected MAEL sequences (*ClustalW* in *MEGA6*). Accession numbers and common name of the organism are in the sequence labels. The alignment was adjusted to reflect secondary structural characteristics observed in NMR structure of the human MAEL HMG-box (PDB ID: 2cto). The secondary structural elements (α-helices and loops) and completely conserved residues are shown on the right. The phylogenic classes for which multiple specie sequences were present are described next to the labels (mammals, reptiles, birds, insects). Other classes include: Amphibia – Frog, Echinoidea – Purple sea urchin, Hydrozoa – Fresh water (FW) polyp. Residues are colored according to the Taylor color scheme to provide contrast to groups of residues (*JalView*).

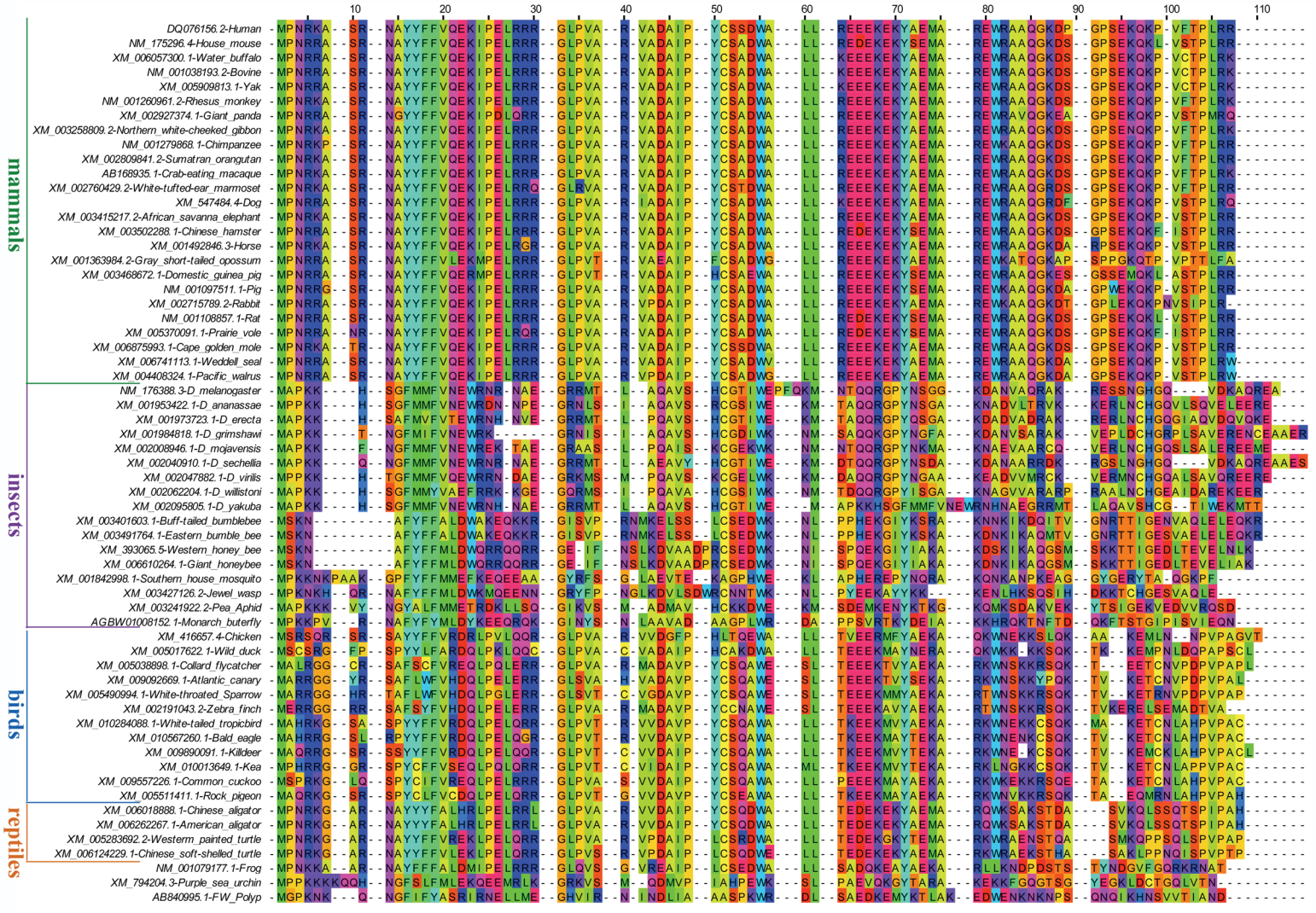
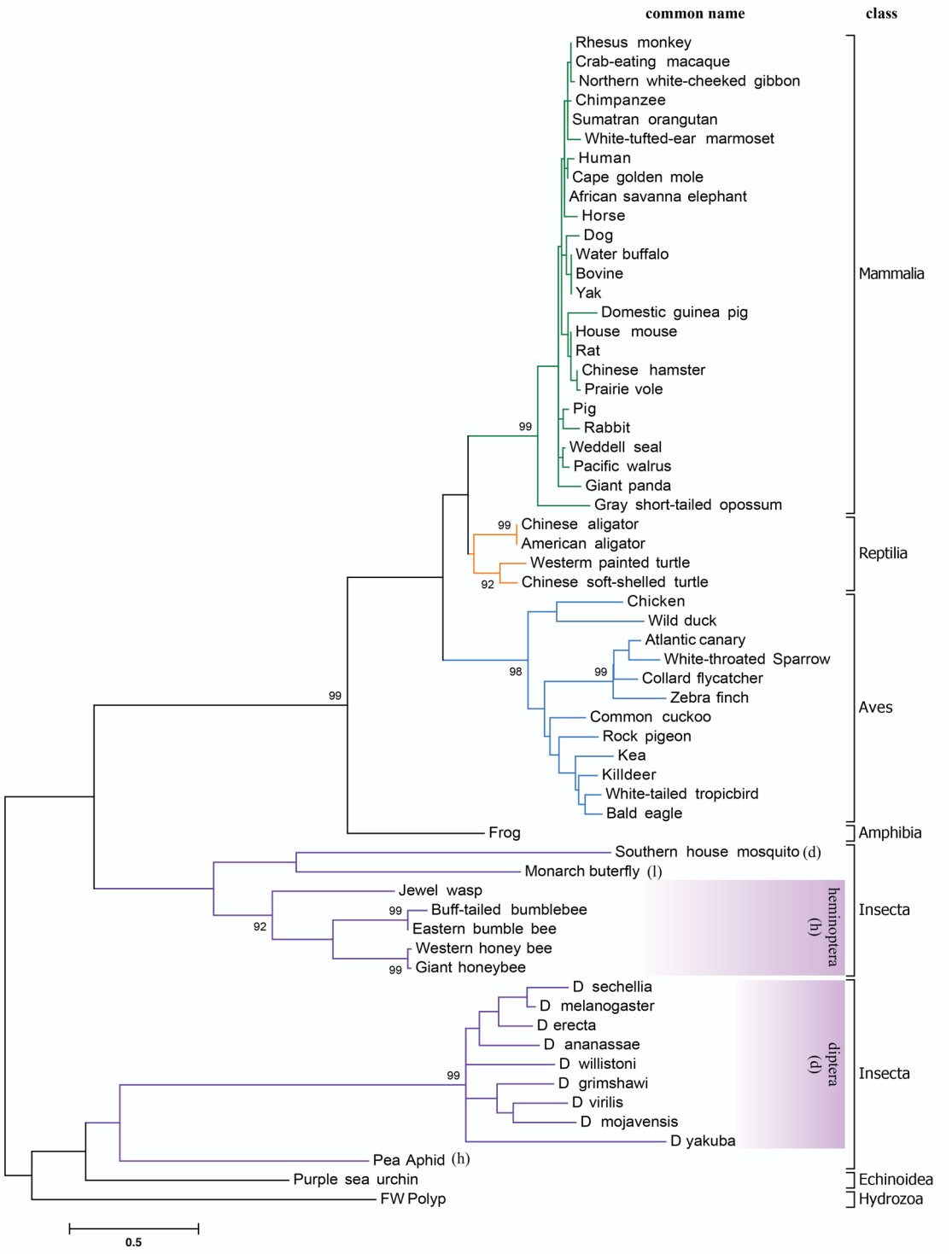

mammals
insects
birds
reptiles
DQ076156.2-Human
NM_175296.4-House_mouse
XM_006057300.1-Water_buffalo
NM_001038193.2-Bovine
XM_005909813.1-Yak
NM_001260961.2-Rhesus_monkey
XM_002927374.1-Giant_panda
XM_003258809.2-Northern_white-cheeked_gibbon
NM_001279868.1-Chimpanzee
XM_002809841.2-Sumatran_orangutan
AB168935.1-Crab-eating_macaque
XM_002760429.2-White-tufted-ear_marmoset
XM_547484.4-Dog
XM_003415217.2-African_savanna_elephant
XM_003502288.1-Chinese_hamster
XM_001492846.3-Horse
XM_001363984.2-Gray_short-tailed_opossum
XM_003468672.1-Domestic_guinea_pig
NM_001097511.1-Pig
XM_002715789.2-Rabbit
NM_001108857.1-Rat
XM_005370091.1-Prairie_vole
XM_006875993.1-Cape_golden_mole
XM_006741113.1-Weddell_seal
XM_004408324.1-Pacific_walrus
NM_176388.3-D_melanogaster
XM_001953422.1-D_ananassae
XM_001973723.1-D_erecta
XM_001984818.1-D_grimshawi
XM_002008946.1-D_mojavensis
XM_002040910.1-D_sechellia
XM_002047882.1-D_virilis
XM_002062204.1-D_willistoni
XM_002095805.1-D_yakuba
XM_003401603.1-Buff-tailed_bumblebee
XM_003491764.1-Eastern_bumble_bee
XM_393065.5-Western_honey_bee
XM_006610264.1-Giant_honeybee
XM_001842998.1-Southern_house_mosquito
XM_003427126.2-Jewel_wasp
XM_003241922.2-Pea_Aphid
AGBW01008152.1-Monarch_buterfly
XM_416657.4-Chicken
XM_005017622.1-Wild_duck
XM_005038898.1-Collard_flycatcher
XM_009092669.1-Atlantic_canary
XM_005490994.1-White-throated_Sparrow
XM_002191043.2-Zebra_finch
XM_010284088.1-White-tailed_tropicbird
XM_010567260.1-Bald_eagle
XM_009890091.1-Killdeer
XM_010013649.1-Kea
XM_009557226.1-Common_cuckoo
XM_005511411.1-Rock_pigeon
XM_006018888.1-Chinese_aligator
XM_006262267.1-American_aligator
XM_005283692.2-Westerm_painted_turtle
XM_006124229.1-Chinese_soft-shelled_turtle
NM_001079177.1-Frog
XM_794204.3-Purple_sea_urchin
AB840995.1-FW_Polyp
2° structure
N-term.
h1
L1
h2
h3
C-term.
100% conserved

Figure 9: Phylogenetic relationship of MAEL HMG-boxes

Phylogenetic analysis of MAEL HMG-box domains peptide sequence (residues 1-86). The tree compares HMG-box domains (73 positions analyzed) to infer their relationships using maximum likelihood (ML) method (*MEGA6*). The values next to the branches describe the percentage of trees where associated sequences group together (n=1000). Only values greater than 90 are shown. The branch length scale is in residue substitutions per site. Shortcuts used for orders within insect class are as follows: l - lepidoptera, d - diptera, h - heminoptera.


common name
class
Rhesus monkey
Crab-eating macaque
Northern white-cheeked gibbon
Chimpanzee
Sumatran orangutan
White-tufted-ear marmoset
Human
Cape golden mole
African savanna elephant
Horse
Dog
Water buffalo
Bovine
Yak
Domestic guinea pig
House mouse
Rat
Chinese hamster
Prairie vole
Pig
Rabbit
Weddell seal
Pacific walrus
Giant panda
Gray short-tailed opossum
Mammalia
Chinese aligator
American aligator
Westerm painted turtle
Chinese soft-shelled turtle
Reptilia
Chicken
Wild duck
Atlantic canary
White-throated Sparrow
Collard flycatcher
Zebra finch
Common cuckoo
Rock pigeon
Kea
Killdeer
White-tailed tropicbird
Bald eagle
Aves
Frog
Amphibia
Southern house mosquito (d)
Monarch buterfly (l)
Jewel wasp
Buff-tailed bumblebee
Eastern bumble bee
Western honey bee
Giant honeybee
heminiptera (h)
Insecta
D sechellia
D melanogaster
D erecta
D ananassae
D willistoni
D grimshawi
D virilis
D mojavensis
D yakuba
diptera (d)
Insecta
Pea Aphid (h)
Purple sea urchin
Echinoidea
FW Polyp
Hydrozoa
99
99
98
99
99
92
99
99
0.5

Section summary

The phylogenetic analysis shows that MAEL HMG-boxes are an old domain with an overall high conservation. Even though the obtained phylogenetic tree does not reflect real time, but relative distance, it is reassuring to see that the tree branches approximately describe relative speciation times of different animal classes. The fresh water (FW) polyp, sea urchin and fruit fly MAEL HMG-boxes belong to the oldest animal classes that diverged from mammals close to one billion years ago, whereas reptile and bird MAEL HMG-boxes have diverged from mammals roughly 300 million years ago [287, 288]. Canonical HMG-box domains are old domains which first appeared over one billion years ago [244], which correspond to the divergence times of the oldest MAEL HMG-box sequences in our data set (fruit fly and fresh water polyp), implying that this domain may have emerged or split from canonical HMG-box domains around the same time.

Another interesting observation relates to the fact that the longest speciation event within the fruit fly took about 50 million years, while the mouse and human are separated by about 92 million years of evolution [287, 288]. The branches in the tree within the *Drosophilae* family of insect class are much longer than those within the entire mammalian class, suggesting that this domain acquired many more changes in a shorter period of time in insects while significantly fewer changes in mammals in an approximately twice-longer period of time (Figure 9). The differences in rate of residue substitution between the insect and mammalian classes may be of functional significance and a subject of a very interesting study in the future. This inquiry will focus on the mouse HMG-box that is near identical to its human counterpart.

Unique Sequence and Structural Features of MAEL HMG-box

To gain functional insight, I have compared human, mouse, and fruit-fly MAEL HMG-box sequences to a subset of sequences of canonical sequence-specific (SS) and non-sequence-specific (NSS) HMG-box domains. The sequence, structural features, and functions of canonical HMG-box domains have been well described and therefore may offer clues about the function of MAEL HMG-box.

Using multiple-sequence alignment of candidate sequences I constructed a phylogenetic tree (*MEGA6*) [246] that relates amino acid sequences of selected HMG-box domains to each other (Figure 10A). Domains of individual HMG-boxes are grouped together in clusters according to their functional characteristics. The HMG-box domains of SRY and SRY-related HMG-box (SOX) proteins formed a branch in agreement with their classification as SS binders [220, 289]. Grouping of these domains together is also in agreement with them being found as a single HMG-box within their respective proteins [214]. The di-domain HMG-box-containing proteins, high-mobility group proteins (HMGBs) 1-3 and Dsp1, which are NSS binders, formed two neighboring branches corresponding to domains A and B (Figure 10A). The branches of the domain A and B are closer to each other than to SS or MAEL HMG-boxes, likely owing the fact that these two domains are within the same peptide and have evolved to accomplish protein-specific functions in concert [290]. On the other hand, the appearance of domains A and B on separate branches reflects on variable sequence characteristics that underlie their observed functional differences [237, 291, 292]. The MAEL HMG-box domains form a separate branch on the tree and are most closely related to the domain A of NSS HMG-box proteins.

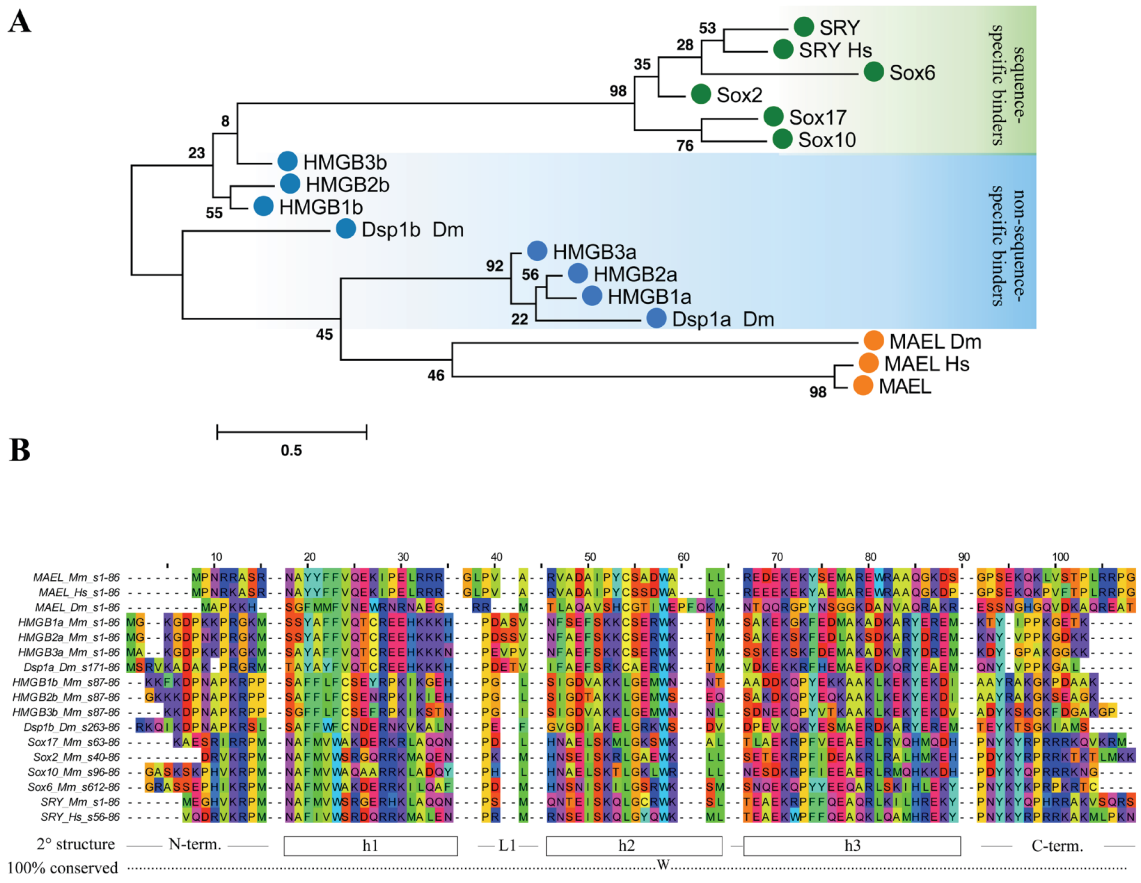


Figure 10: Phylogenetic relationships of MAEL and canonical HMG-boxes

A) Phylogenetic analysis using the maximum likelihood method comparing canonical HMG-boxes with MAEL HMG-boxes (73 positions analyzed). Mouse sequences are shown unless otherwise noted (Dm: *Drosophila melanogaster,* Hs: *Homo sapiens*). Values describe percentage of trees where associated sequences group together (n=1000). The branch length scale is in number of substitutions per site. B) Multiple amino acid sequence alignment of candidate HMG-box domains. The residues were pseudo-colored according to the Taylor color scheme (*JalView*) to contrast side chain chemical characteristics. Completely conserved secondary structure elements and residues are shown below the alignment: h – α-helix, L – loop.

The relative distance to SS HMG-boxes and proximity to NSS HMG-boxes imply that MAEL HMG-box has residues that may provide means for NSS binding. Furthermore, formation of a separate branch on the tree by MAEL HMG-boxes also suggests that these domains have additional sequence characteristics that set them apart (Figure 10A). Based on the branch length, the fruit fly Mael HMG-box domain is a distant relative of its mammalian counterparts, agreeing with my earlier observations that mammalian and insect classes of MAEL HMG-box domains are separated by approximately one billion years of evolution. However, because fruit fly domain groups with other MAEL HMG-boxes, as opposed to grouping with canonical SS and NSS HMG-boxes, it is likely that this domain has sequence or structural features in common with mammalian domains, suggesting a common ancestor domain (Figure 10A).

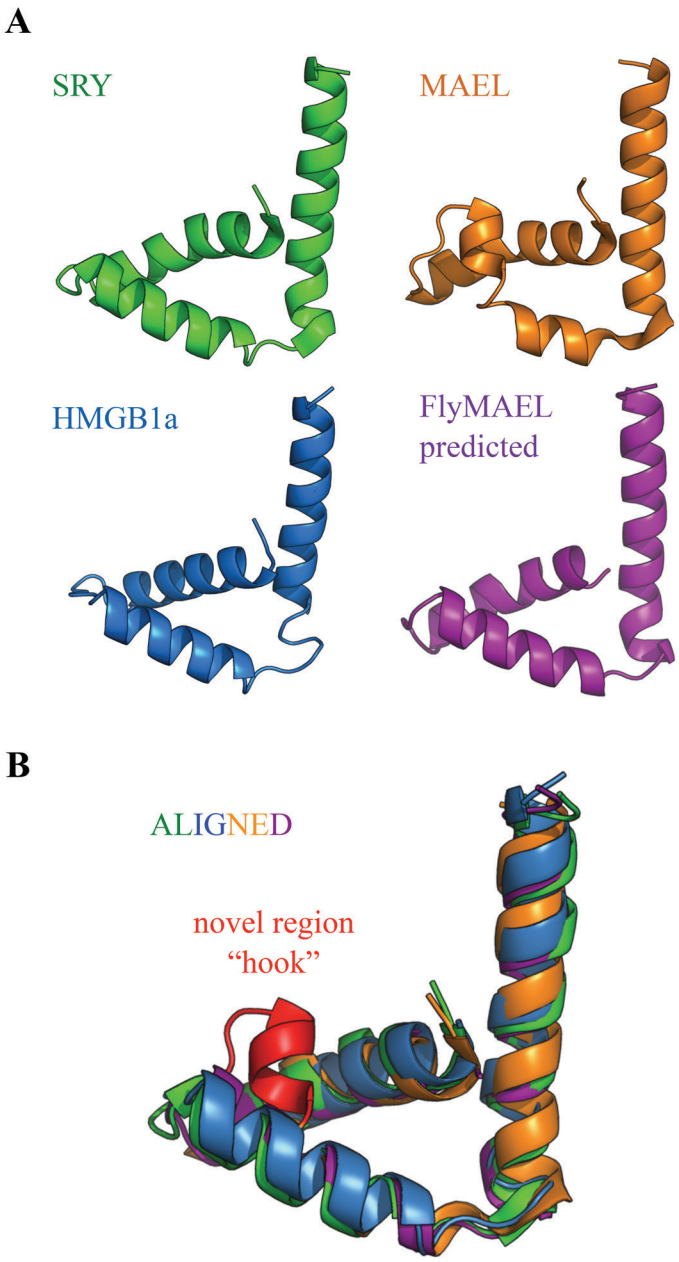
Inspection of the sequence alignment reveals a number of sequence features underlying the observed phylogenetic clustering (Figure 10B). First, the distribution of the charged residues within MAEL HMG-boxes is different. In the mammalian MAEL HMG-boxes, the loop (L1) connecting α-helix-1 and 2 are hydrophobic and don't contain any charged residues. However, charged residues are present in the domain A of HMGBs and SS HMG-boxes (Figure 10B). Charged residues in SS and NSS domains appear to be alternating throughout α-helix-1 and 2, but in the mammalian MAEL HMG-boxes positive residues are concentrated on both sides of the L1. The distribution of the charged residues can be indicative of H-bonding potential that together with hydrophobic regions can provide the biochemical basis for strong interactions with an appropriate substrate. These features vary in the fruit fly Mael HMG-box where the residues in α-helix-1 are distributed in a fashion that resembles both SS and NSS domains, and unlike its

mammalian homologues, the positive residues are within its L1 region (Figure 10B). However, similarly to the mammalian counterparts, these residues are consecutive arginines and are not observed in the corresponding regions of the SS and NSS domains. This suggests that while the position of charged residues in the fruit fly has diverged, their chemical identity is conserved and therefore the L1 region and its surroundings may be functionally analogous to their mammalian versions.

Phylogenetic and sequence analysis highlighted features unique to MAEL HMG-boxes as well as revealed MAEL HMG-box's relatively close relationship to NSS HMG-boxes. Next, I performed a structural comparison between SS, NSS and MAEL HMG-box domains. To do this, I used previously determined structures of SRY HMG-box (PDB ID: 1j46, SS), HMGB1a box (PDB ID: 1ckt, NSS), human MAEL HMG-box (PDB ID: 2cto) and computationally predicted (*Robetta*) *Drosophila mealnogaster* Mael HMG-box. Structurally, all HMG-box domains have a characteristic L-shaped fold composed of three helices (Figure 3, Figure 11) [293]. Similarly to the HMG-box domains of SRY and HMGB1, both mouse and fruit fly domains also share this fold (Figure 11A). However, mouse MAEL HMG-box has acquired a bend in α-helix-2 that is immediately evident in the structural alignment (Figure 11B). The α-helix-2 bends approximately 45 degrees from the other aligned α-helices, resembling a "hook". The fruit fly Mael HMG-box does not have this bend in the same α-helix. Its tertiary structure was modeled on the structure of the HMG-box of Sox2 (PDB ID: 1gt0 chain D), which also features a straight second α-helix. Some sequence features of the fruit fly domain resemble that of the Sox2 HMG-box; however, the phylogenetic analysis shows that this domain is as distinct as other MAEL HMG-boxes are (Figure 10).

Figure 11: MAEL HMG-boxes structural comparison

A) Experimentally determined as well as *in silico* predicted (*Drosophila melanogaster* (fly) MAEL HMG-box) structures of candidate HMG-box proteins. SRY: 1j46 - sequence specific binding; HMGB1a: 1ckt - structure specific binding; mouse MAEL: 2cto - unknown binding; flyMAEL - unknown binding. B) MAEL HMG-box domains have a conserved canonical L-shape. Mouse MAEL has a bend in α-helix-2, creating a novel region termed “hook” (red). This feature is absent in the predicted structure of flyMAEL domain.

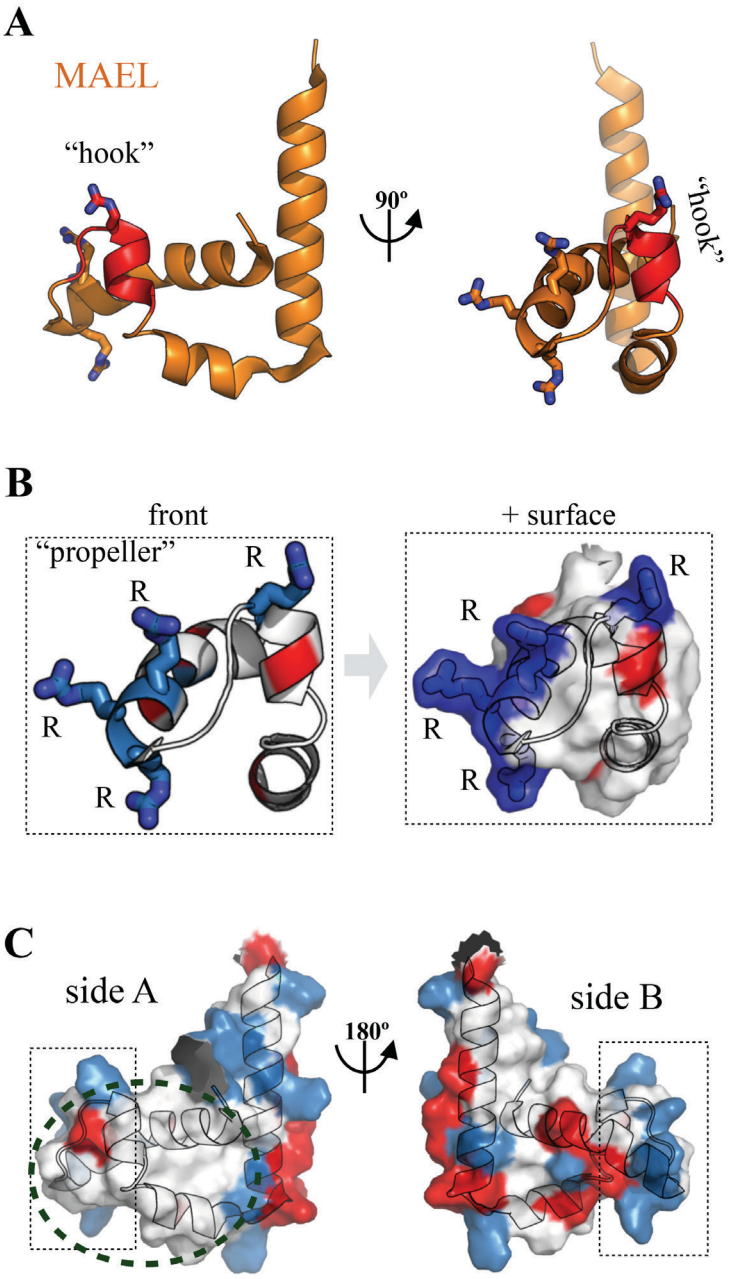

A
SRY
MAEL
HMGB1a
FlyMAEL
predicted
B
ALIGNED
novel region
“hook”

Therefore, absence of the "hook" seen in the mouse HMG-box may be an artifact stemming from *Robetta* modeling the fruit fly domain onto the structure of Sox2. The possibility that a structural feature analogous to the mammalian "hook" exists in the fruit fly domain cannot be excluded, and its presence would be best confirmed by an experimental determination of the fruit fly MAEL HMG-box structure that may provide additional insights into the role of domains' residues and overall function. Nevertheless, the above structural alignment reveals that mammalian MAEL HMG-boxes have a novel feature ("hook") that, given its intriguing location, is likely to have functional consequences [293].

My analysis of HMG-box sequences has shown that the mouse MAEL HMG-box has several positive residues proximal to L1 region, and uniquely these residues are arginines (Figure 10B). In context of the tertiary structure, these residues group closely to each other, in near proximity to the previously described "hook" region (Figure 12A). The "hook" region contains a single arginine residue, while three additional consecutive arginines are located at the C-terminus of α-helix-1. Their side-chains protrude outward in different directions spanning approximately a ¾ turn (270 degrees), which gives this region the appearance of a "propeller" (Figure 12B). Together, the "hook" and "propeller" form a novel region containing four positively charged arginine residues in close proximity to each other. Charged residues are also present within canonical HMG-boxes, however, their distribution and residue type differ significantly (Figure 10B). They are found internally within α-helix-1 in SS binders, and are not arginines in NSS binders (Figure 10B).

Figure 12: Unique features of mouse MAEL HMG-box

A) Location of the arginine containing "hook" and "propeller" regions within mouse MAEL HMG-box tertiary structure. Arginines (R) are depicted as sticks. B) Close up of the "propeller" and "hook" regions demonstrating outwards side chain protrusion. Sticks and surface presentation are colored according to charge (red – negative, acidic; blue – positive, basic) C) Distributions of positively and negatively charged residues on MAEL HMG-box surface. Uncharged (white) patch of residue on side A is highlighted (dashed green line).


A
MAEL
"hook"
90°
"hook"
B
front
"propeller"
R
R
R
R
+ surface
R
R
R
R
C
side A
180°
side B

Arginine residues have been previously described to play prominent roles in nucleic acid interactions [294-296], therefore the four arginine residues described above may be important for function of MAEL HMG-box. Even though an analogous structural region within predicted that the fruit fly Mael HMG-box is absent, possibly because of an artifact of the software-modeling algorithm (Figure 11A), this domain still has two arginine residues in the nearby region, implying evolutionary conservation (Figure 10B).

To examine the distribution of charged residues within the mouse MAEL HMG-box, I have colored positive and negative side-chains (Figure 12C). Surface rendering shows that the novel arginine-rich region is bulky and positively charged. However, this representation also shows a difference in the distribution of charged residue between side A and B of the MAEL HMG-box. While positive and negative residues are distributed on side B, side A has a single patch of negative residues, few positive residues and a large uncharged region (Figure 12C). The uncharged region seen here corresponds to the region, which interfaces with nucleic acids in SS and NSS canonical domains [219, 221]. The uncharged portions of proteins often contribute to binding through hydrophobic burial [297]. Therefore, together with the charged "hook" and "propeller", the hydrophobic region on side A may contribute to the additional and necessary binding surface required for formation of a strong complex with the appropriate substrate.

Section summary

The phylogenetic classification of the fruit fly Mael HMG-box suggests that just like its mammalian counterpart; this domain is distinct from canonical HMG-boxes but also evolved features distinct from its mammalian homologues, which may perhaps reflect

species-specific specialization of protein function. The mouse and human MAEL HMG-boxes are highly similar to each other with only few biochemically conservative residue substitutions, which is unlikely to affect domain function. However, sequence conservation with fruit fly homologues is much weaker, likely owing to the split of the species about one billion years ago. Fruit fly domains have also evolved at a much faster pace than mammalian ones, which may be reflective of different functional requirements driving their evolution. Nevertheless, both fruit fly and mouse MAEL HMG-box domains have characteristics that set them apart from canonical HMG-box domains. Both domains have arginine residues in the region connecting their first two α-helices, while the SS and NSS domains either do not have arginines in the equivalent region or they are located elsewhere. In the mouse MAEL HMG-box, arginines are located within the region that is important for function of the canonical HMG-box domains. This distinct distribution of arginine residues gives rise to unique structural arrangements, "hook" and "propeller." The domain's charged residue distribution, together with the novel arginine-rich arrangements, are likely to be of functional significance.

Interrogation of MAEL HMG-box Binding to Nucleic Acids

My previous sequence and structural analysis of MAEL-HMG-box suggests that this domain is distinct from SS and NSS HMG-boxes. The positive residues, uniquely arginines, within MAEL HMG-box are enriched at α-helical termini on the short arm of the domain surrounding the uncharged L1 region. As a result, this gives rise to novel structural regions, "hook" and "propeller", both of which are likely to influence this domain's binding to nucleic acid substrates.

In order to evaluate the biochemical activity of MAEL HMG-box domains, I have cloned the HMG-box-containing regions into GST overexpression vector and purified HMG-box domains of SRY (SS), HMGB1a (NSS), mouse MAEL and fruit-fly MAEL from bacteria. The optimized two-step purification scheme is described in detail in the Material and Methods chapter (Figure 13). To confirm that purified proteins are folded into HMG-box fold, I measured ellipticity of the peptide backbone by circular dichroism (CD) in the far-UV range (195-265 nm). Helical proteins exhibit negative ellipticity in this range with two minima at 208 and 222 nm wavelengths [251]. In accordance with these criteria all of the purified proteins showed appropriate negative ellipticity and minima, confirming that they are highly helical and take on canonical HMG-box fold (Figure 14).

I then used gel-shift assays to evaluate the ability of the different purified domains to bind to various nucleic acid substrates. Gel-shift assays are a well-established method used for identification and characterization of protein-nucleic acid interactions [255, 298, 299]. They have been used extensively for descriptions of protein binding to both DNA and RNA [225, 235, 237].

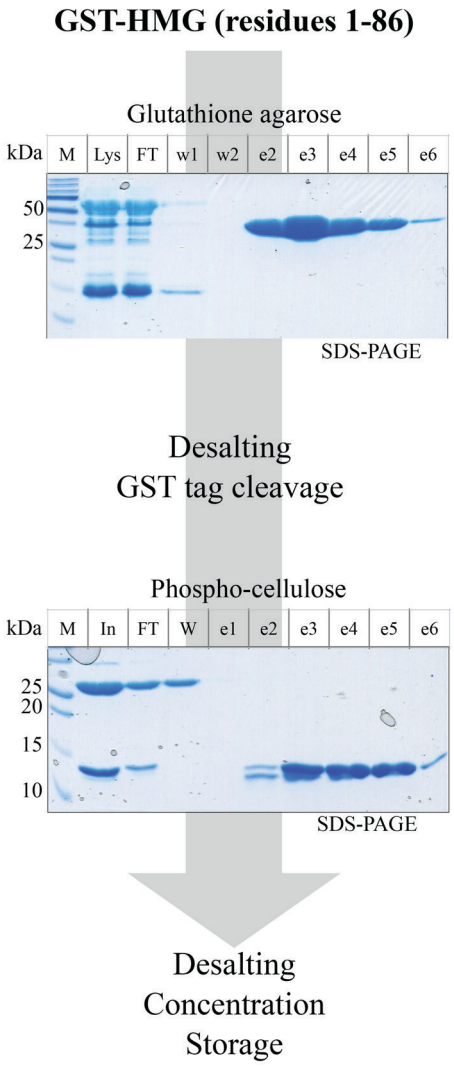


Figure 13: HMG-box purification scheme

First 86 residues of each HMG box were expressed as GST-fusion proteins in bacteria. These were purified on glutathione agarose columns and desalted. The GST tag was then cleaved and the protein purified by ion exchange on phospho-cellulose columns. The final eluate was desalted, concentrated and appropriately stored. Lys-lysate, In-input, FT-flow-through, W-wash, E-eluate.

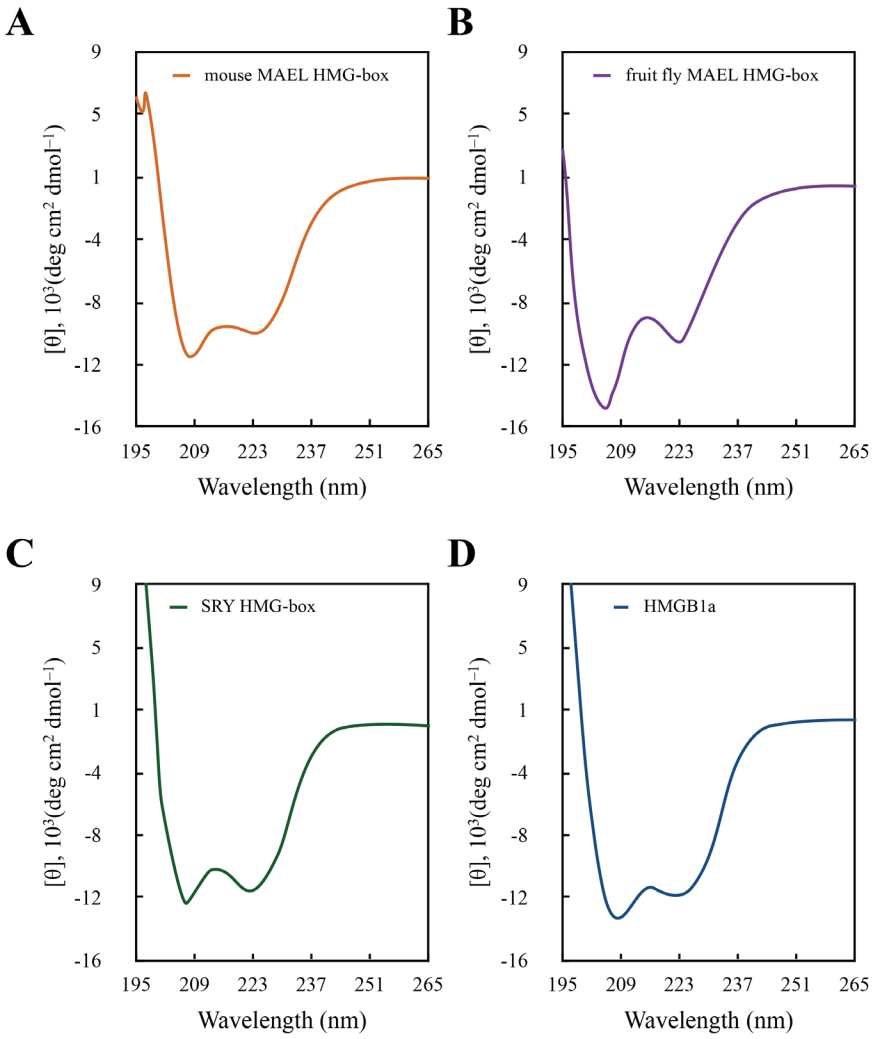


Figure 14: Circular dichroism (CD) of purified HMG-box domains

A) The mouse MAEL HMG-box domain. B) The fruit fly Mael HMG-box. C) Sequence-specific SRY HMG-box. D) Non-sequence-specific HMGB1a. Shown far-UV spectra have negative ellipticity in this range with two minima at 208 and 222 nm typical of highly α-helical proteins [251].

By titrating increasing amounts of protein to a known concentration of radioactively (γ-P^{32}) labeled ('hot') nucleic acid and quantifying the free and bound ('shifted') fraction, I calculated the apparent dissociation constants (K_D) for different HMG-box domains where applicable. In cases where binding was strong and most of the substrate was shifted, I used competition assays in which both the protein concentration (sufficient to achieve 60-90% of total binding) along with the 'hot' nucleic acid concentration were kept constant, and instead an unlabeled ('cold') nucleic acid substrate was titrated in. This allowed me to calculate the competitor dissociation constant (K_C), which is directly related to the K_D. When considered together, these two constants provide a more reliable measurement of binding strength [255].

MAEL HMG-box Does Not Bind dsDNA

First I queried the interaction of the mouse MAEL HMG-box with randomly selected 26-nt long single-stranded (ss) DNA and did not observe any binding (Figure 13, Table 2). The proteins with single HMG-box domains are best known for binding to dsDNA, preferentially recognizing specific nucleotide motifs [214]. For example, transcription factor SRY HMG-box domain binds and bends AT-rich dsDNA featuring an AACAAN consensus sequence [220, 289, 300]. I designed a 16-base pair long dsDNA substrate possessing the SRY census motif and tested it with all purified HMG-boxes. The SRY HMG-box bound to this substrate well, forming a single complex at the lowest and shifting all of the substrate at the highest protein concentration tested (Figure 16A). The estimated K_D was ~16.3 nM, reflecting the strength of this interaction (Figure 16 A`).

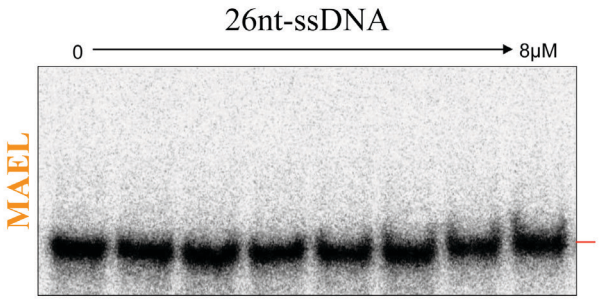
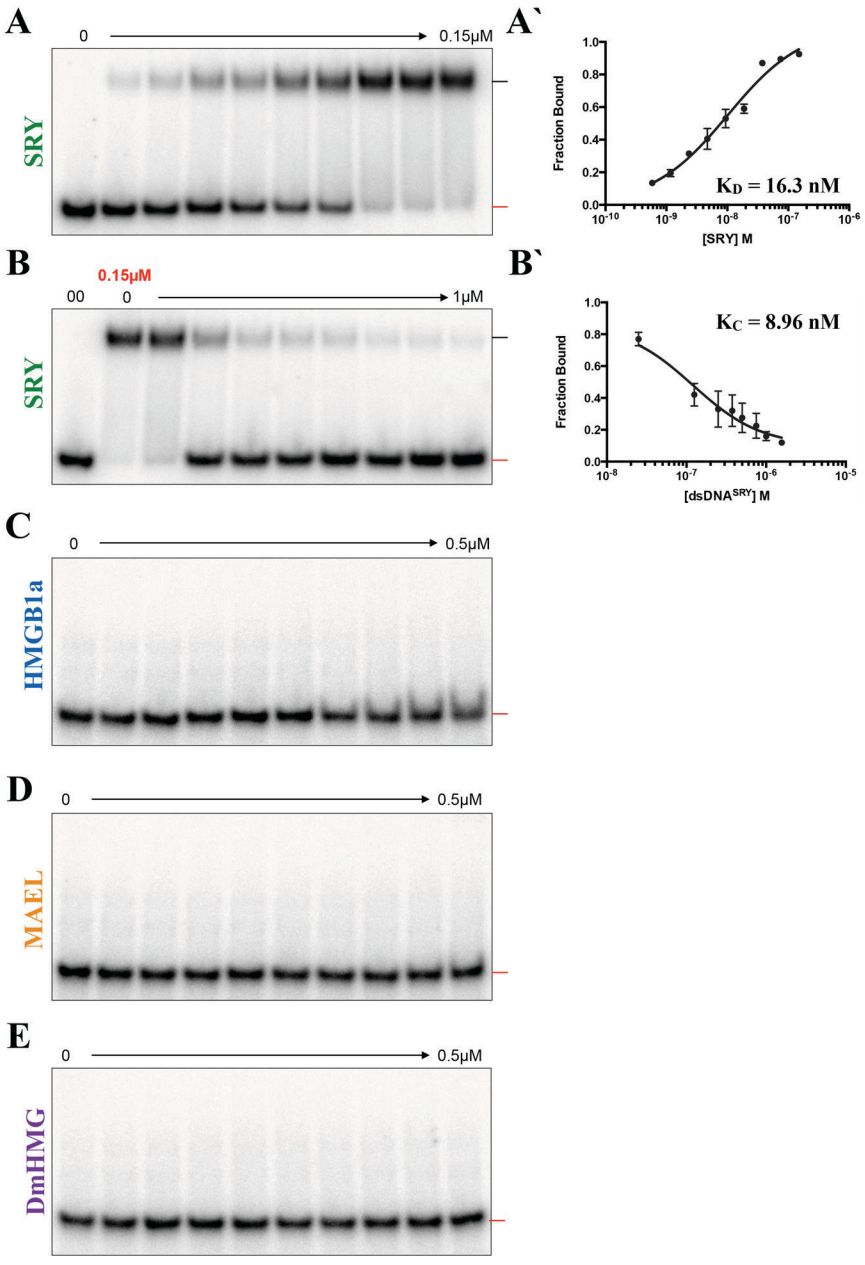


Figure 15: Single stranded (ss) DNA interactions

The mouse MAEL HMG-box does not bind to single-stranded DNA.

Figure 16: Double stranded (ds) DNA interactions

A) Titration of the SRY HMG-box binds to 16 base pair long dsDNA containing its consensus sequence motif (AACAAT). A`) Binding affinity curve calculated from replicate titration experiments shows that the SRY HMG-box binds to this substrate strongly. Dissociation constant (K_D) is 16.3 nM. B) Competition assay using 'cold' dsDNA leads to titration of the protein away from labeled substrate. The 00 lane contains only 'hot' substrate, and the red value indicates the amount of protein in all the following lanes. B`) Competitor binding affinity curve confirms high affinity of and specificity of SRY HMG-box interaction with dsDNA containing its preferred sequence. Competitor dissociation constant (K_C) is 8.96 nM. Titration of C) HMGB1a, D) mouse MAEL HMG-box, or E) fruit fly Mael HMG-box domain to same substrate does not show any binding. Lines on the side of gels indicate free (**red**) and bound (**black**) substrates.


A
0
0.15μM
SRY
A`
Fraction Bound
1.0
0.8
0.6
0.4
0.2
0.0
10^{-10}
10^{-9}
10^{-8}
10^{-7}
10^{-6}
[SRY] M
K_D = 16.3 nM
B
0.15μM
00
0
1μM
SRY
B`
Fraction Bound
K_C = 8.96 nM
10^{-8}
10^{-7}
10^{-6}
10^{-5}
[dsDNA^{SRY}] M
C
0
0.5μM
HMGB1a
D
0
0.5μM
MAEL
E
0
0.5μM
DmHMG

The 'cold' competition assay and calculated K_C (~8.96 nM) confirmed the strength and specificity of this interaction (Figure 16B, B`). The average binding affinity in my experiments (~12 nM) is close to the values reported in literature (~20 nM) [301]. By contrast, the HMGB1a, mouse and fruit fly MAEL HMG-box domains did not bind to the same substrate (Figure 16C-E). I tested mouse and fruit fly Mael HMG-boxes with an extensive series of dsDNA substrates (Table 3), varying the length and base compositions as well as including symmetrical cytosine methylation. Surprisingly, neither HMG-box domain formed a complex with any of the dsDNA substrates. I used specific DNA sequences described in a study of fruit fly Mael in which its DNA-binding was investigated by chromatin immunoprecipitation [166]. To my surprise, I did not observe any binding to the reported sequences in my experiments (Figure 16D, E, Table 3). My sequence and predicted structure analyses suggested that fruit fly Mael HMG-box has diverged from its mammalian counterparts (Figure 9, 10B). Despite this, it also fails to bind to the dsDNA substrates tested here, just like its mouse counterpart (Figure 16D, E). Additionally, I was not able to identify structural regions homologous to "hook" and "propeller" in fruit fly Mael HMG-box, but it does have two arginine residues that, like arginines in mammalian domains, may give rise to similar structural features (Figure 10B).

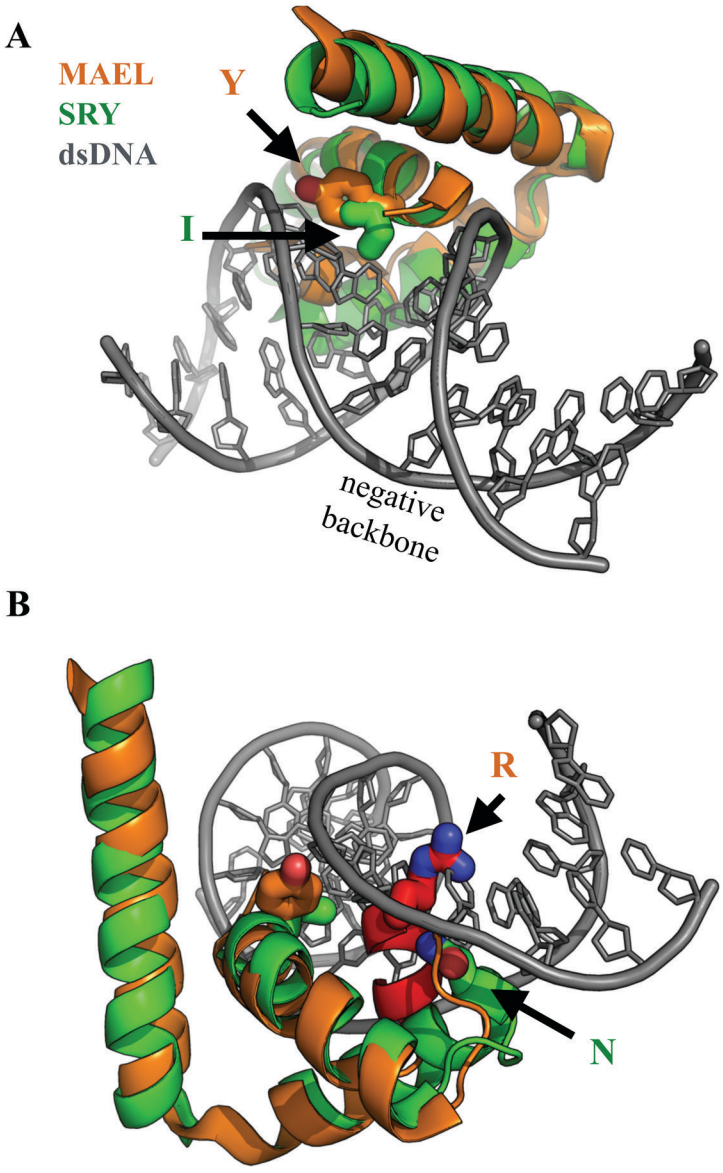
Section summary

In accordance with previous observations, SRY HMG-box binds strongly to dsDNA containing its recognition sequence motif [289]. It does so by first intercalating an isoleucine (Ile) residue from the N-terminal end of α-helix-1 between bases in the AT-

rich minor groove region of the dsDNA, which results in perturbation of the helical geometry. Next, bent DNA helix accommodates residues along α-helix-1 and 2 within the widened DNA minor groove. The complex is stabilized through multiple DNA base and backbone contacts with the different α-helices and termini of the SRY HMG-box (Figure 4A) [219]. HMGB1a cannot bind to the same DNA because it lacks sequence and structural features required for DNA bending. It has alanine (Ala) at the equivalent position, which has a short side chain and cannot intercalate between bases and bend the DNA helix. Instead, the aromatic side-chain of phenylalanine (Phe) at the beginning of the α-helix-2 gets accommodated in the pocket formed by cis-Pt modification of dsDNA (Figure 4B) [221, 223, 293]. The mammalian MAEL HMG-box has tyrosine (Tyr) in the N-terminal region of α-helix-1, whereas SRY HMG-box has an Ile residue (Figure 17A). When modeled with dsDNA, tyrosine residue is positioned so that it appears to be facing the phosphate backbone of the DNA. Not only does its position preclude intercalation between bases, but it also may cause electrostatic repulsion of the negatively charged DNA backbone with the Tyr hydroxyl group (Figure 17A). Furthermore, the novel "hook" region of α-helix-2 of mammalian MAEL HMG-box cannot be accommodated within the dsDNA grooves (Figure 17B). Therefore, both of these regions likely contribute to making interactions with the canonical B-type helix of dsDNA unfavorable. Additionally, the "hook" and "propeller" are readily detected in NMR structure of the human domain but are not identified by the structural prediction algorithm for the fruit fly Mael HMG box. A presence of two arginine residues in the fruit fly suggests that similar regions may nevertheless be present, which is also supported by the failure of both of these domains to bind to various dsDNA substrates.

Figure 17: Structural model of dsDNA binding.

A) Structural alignment of the MAEL HMG-box to solved structure of the SRY HMG-box bound to dsDNA containing its recognition sequence ($dsDNA^{SRY}$). The SRY HMG-box intercalating residue (Ile - green spheres) points toward the center of the dsDNA helix. MAEL HMG-box has two tyrosines (Tyr – orange spheres) in equivalent region, both of which point towards DNA phosphate backbone. B) The mouse MAEL HMG-box has a bend in its second α-helix that forms what looks like a "hook". This region protrudes into dsDNA helix, precluding same kind of interaction as the SRY HMG-box. In this model the dsDNA is already bent by the SRY HMG-box; however, this bend is not sufficient to accommodate the "hook" region of the MAEL HMG-box.


A
MAEL
SRY
dsDNA
Y
I
negative
backbone
B
R
N

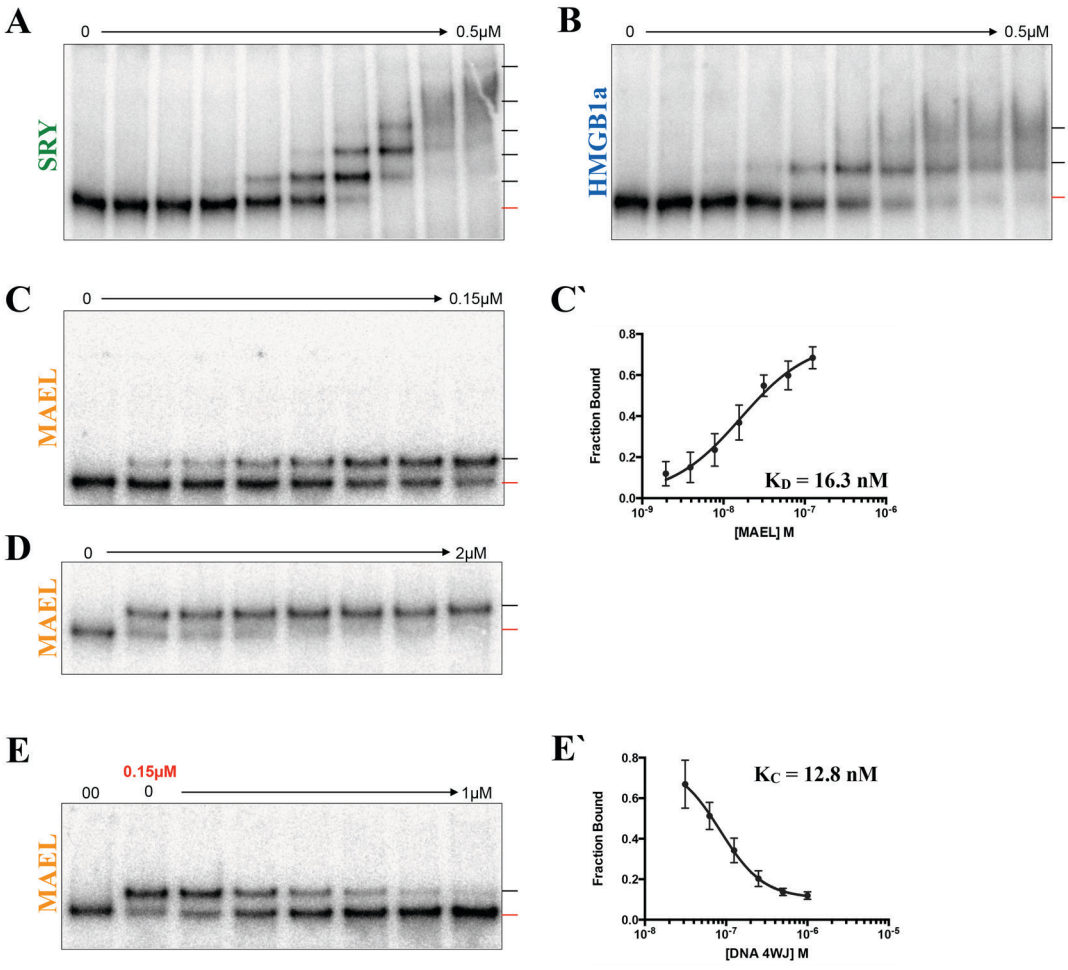
DNA junctions are Strongly Bound by MAEL HMG-box

My previous analysis showed that the MAEL HMG-box domain does not bind to the dsDNA. The most likely reason is the lack of intercalating residue and presence of "hook" and "propeller" that structurally occlude such interactions. Nevertheless, these regions may bind to substrates with non-B-type helical geometry, such as DNA four-way junctions (4WJ). Binding to DNA 4WJs is a commonly observed function of other HMG-box domains [223-226]. These junctions are comprised of four double-stranded arms with a central region where the strands sharply turn and the helical grooves widen; therefore, 4WJs have B-type and non-B-type structural characteristics. The dynamics and structure of DNA 4WJ, also known as Holliday junctions, have been well studied because of their pertinence to recombination and DNA repair [227-231]. The double-stranded arms of DNA 4WJ transition between states where they are extended outward in all directions (open conformation) and the states where they are proximal to each other (closed conformation) in parallel or antiparallel orientation based on whether the individual strands cross each other or not. In the open conformation, the center of the DNA 4WJ has space that can accommodate the binding of (comparatively small) HMG-box domains [225]. The transitions between open and closed conformations are influenced by the abundance of various ions in the solution; however, under physiological conditions, DNA 4WJ are primarily open [228, 302-304].

To determine whether the ability to bind 4WJs is retained by MAEL HMG-box, I tested the mouse domain binding to DNA 4WJ and compared it to binding of SRY HMG-box (SS) and HMGB1a (NSS) to the same substrate. Both SRY HMG-box and HMGB1a readily bind the DNA 4WJ forming multiple complexes (Figure 18A,B),

Figure 18: DNA 4WJ binding

A) Titration of SRY HMG-box to DNA 4WJ. Red lines denote free substrate. Five distinct complexes formed (**black** lines) with consecutive addition of protein. B) Titration of HMGB1a to DNA 4WJ. Two complexes form at the tested protein concentrations. C, D) Titration of mouse MAEL HMG-box to the same substrate results in formation of only single complex even at high concentrations. C`) Binding-affinity curve calculated using replicate experiments shows that mouse MAEL HMG-box binds to DNA 4WJ strongly ($K_D$ = 16.3 nM). D) Competition assay using unlabeled substrate and D`) competitor binding affinity curve confirms strength and specificity of this interaction (K_C = 12.8 nM). The 00 lane contains only 'hot' substrate and the red value indicates amount of protein in all the following lanes.


A
0
0.5μM
SRY
B
0
0.5μM
HMGB1a
C
0
0.15μM
MAEL
C`
Fraction Bound
K_D = 16.3 nM
[MAEL] M
D
0
2μM
MAEL
E
0.15μM
00
0
1μM
MAEL
E`
Fraction Bound
K_C = 12.8 nM
[DNA 4WJ] M

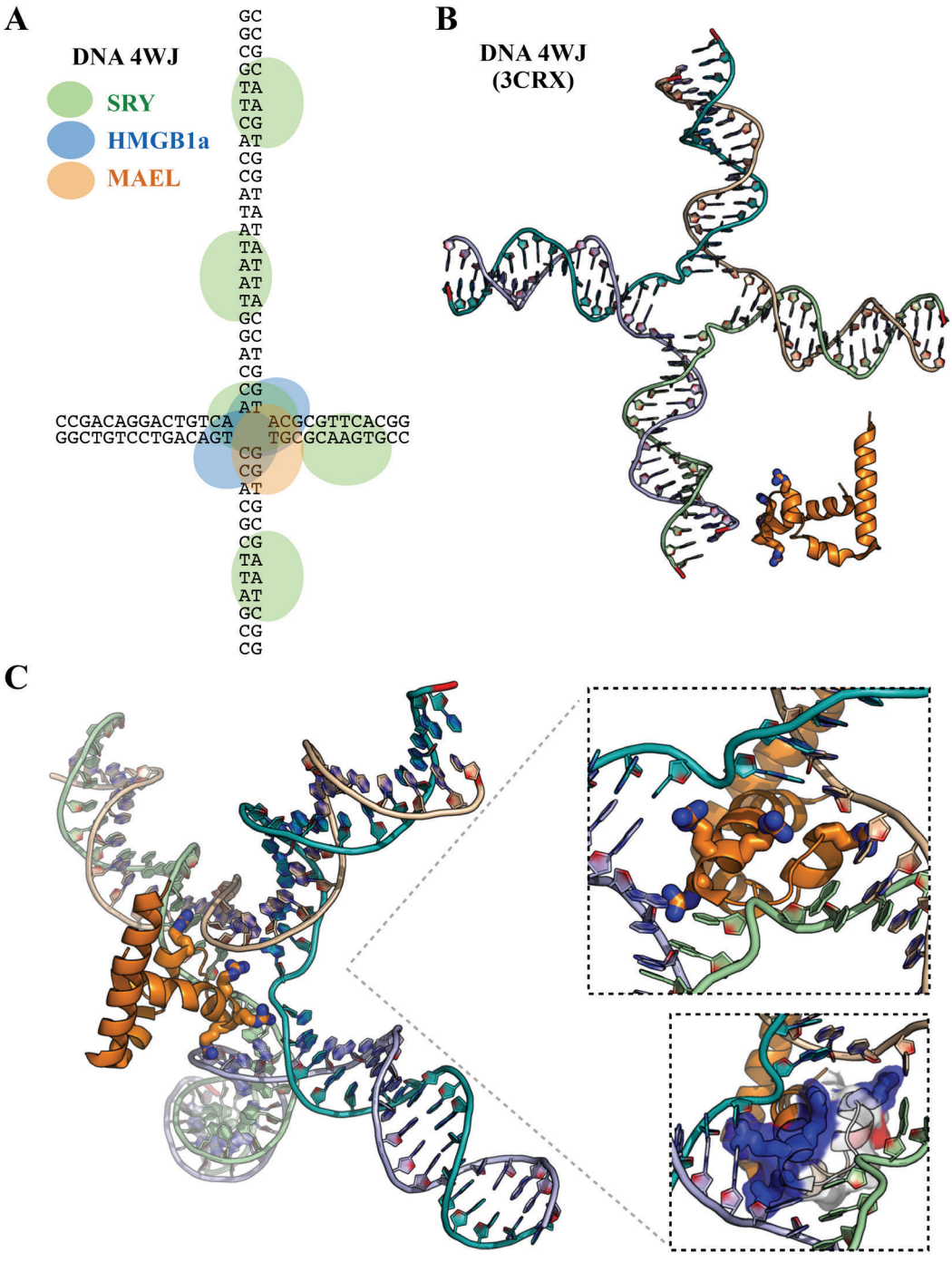
confirming previous observations [225]. Unlike the SRY HMG-box and HMGB1a domains, mouse the MAEL HMG-box domain induced formation of single complex even at high protein concentrations (Figure 16 C, D). This interaction with DNA 4WJ was strong (K_D ~16.3 nM) and specific as indicated by the 'cold' competition assay (K_C ~ 12.8 nM) (Figure 18C`, E`), and the binding affinity was close to the values reported for SRY (K_D ~ 10 nM) [224, 301] and HMGB1 (K_D ~1-10 nM) [305] proteins.

Section summary

SRY HMG-box binds to dsDNA in a sequence-specific manner with preference for AT-rich regions [219]. Its sequence features allow it to bend the DNA and form an optimal site for binding. The DNA 4WJ used here has AT-rich sequences in the double-stranded arms which may be recognized as potential binding sites (Figure 19A). The open center is already perturbed, making it accessible for binding without the need for additional perturbations by the protein domain. Therefore, the five SRY-DNA 4WJ complexes are likely products of binding to the open center and to the AT-rich regions of the double-stranded arms (Figure 18A). The HMGB1a does not bind to dsDNA (Figure 16C, 19A) because it lacks residues required for bending the DNA, facilitating the formation of an open region that can accommodate the Phe residue [221, 306, 307]. However, perturbed non-B-type DNA regions already exist in the center of the DNA 4WJ open conformation. Therefore, the two HMGB1a complexes likely represent binding of two proteins to the open center. Because the Phe residue is at the edge of the domain's short arm (Figure 4), which is not as bulky as homologous region of the mouse MAEL HMG-box, it is likely that two proteins are accommodated at center of the DNA 4WJ (Figure 18B, 19A).

Figure 19: Modeling HMG-box onto DNA 4WJ

A) Proposed model for binding of tested HMG-box proteins DNA 4WJ. Each protein can bind to the center of the junction because it is already pre-bent. The SRY HMG-box can bend dsDNA and therefore is also modeled on AT-rich dsDNA regions approximating its recognition sequence. HMGB1a cannot bend dsDNA; therefore the two observed complexes likely represent two proteins accommodated at perturbed center of 4WJ. Due to bulkiness of its "hook" and "propeller" regions, only one mouse MAEL HMG-box protein can be accommodated at the open center of the DNA 4WJ. Exact sequence of the tested substrate depicted. B) Experimentally solved structures of unrelated DNA 4WJ in open conformation (PDB ID: 3CRX - chains C-F) and human MAEL HMG-box domain (PDB ID: 2cto) represented to scale. C) Suggested mode of MAEL HMG-box binding to open region of DNA 4WJ in (B) based on modeling (*PyMol*). Positive arginine residues (sticks) in the "propeller" and "hook" regions may be involved in multiple DNA base and backbone interactions at the open center of DNA 4WJ. Two close-up views are shown on the right. Each strand of DNA 4WJ is depicted in a different color. The colors of nucleic acid bases and protein side chains represent charge as follows: red – negative, blue – positive.


A
DNA 4WJ
SRY
HMGB1a
MAEL
GC
GC
CG
GC
TA
TA
CG
AT
CG
CG
AT
TA
AT
TA
AT
AT
AT
TA
GC
GC
AT
CG
CG
AT
CCGACAGGACTGTCA
GGCTGTCCTGACAGT
ACGCGTTCACGG
TGCGCAAGTGCC
CG
CG
AT
CG
GC
CG
TA
TA
AT
GC
CG
CG
B
DNA 4WJ
(3CRX)
C

The MAEL HMG-box has structural features distinguishing it from the SRY HMG-box and HMGB1a (Figure 11, 12). It has Tyr residues in place of the intercalating residues and bulky arginine-rich "hook" and "propeller" regions, which precludes interactions with dsDNA (Figure 12, 16D). Therefore, the single complex formed with the DNA 4WJ must correspond to binding to MAEL HMG-box the open center, and only a single protein can be accommodated due to the bulkiness of the novel features of the MAEL HMG-box domain (Figure 18C, D, 19A). Using best-fit modeling of MAEL HMG-box onto the DNA 4WJ, I put forward a single scenario that accommodates my structural and experimental observations (Figure 19B, C). The novel "propeller" and "hook" regions are positioned such that the arginine side chains make multiple contacts with the bases and the backbone of the DNA 4WJ (Figure 19C).

My results show that like other HMG-box domains, the MAEL HMG-box can bind to DNA 4WJ. However, I observed only a single complex whereas SRY HMG-box formed up to five and HMGB1a formed up to two. I believe that this difference is due to the unique structural features of the MAEL HMG-box distinguishing it from the canonical domains. It is not able to bind to dsDNA. As such, it most likely binds to the perturbed center region of the DNA 4WJ. Using *in-silico* modeling I describe how its structural features could support binding to the 4WJ center region. The lack of binding to dsDNA and binding to DNA 4WJ as a single complex taken together suggest that the MAEL HMG-box binding to DNA substrate does not depend on sequence, but rather on the availability of an open DNA structure.

MAEL HMG-box Binds to RNA Molecules

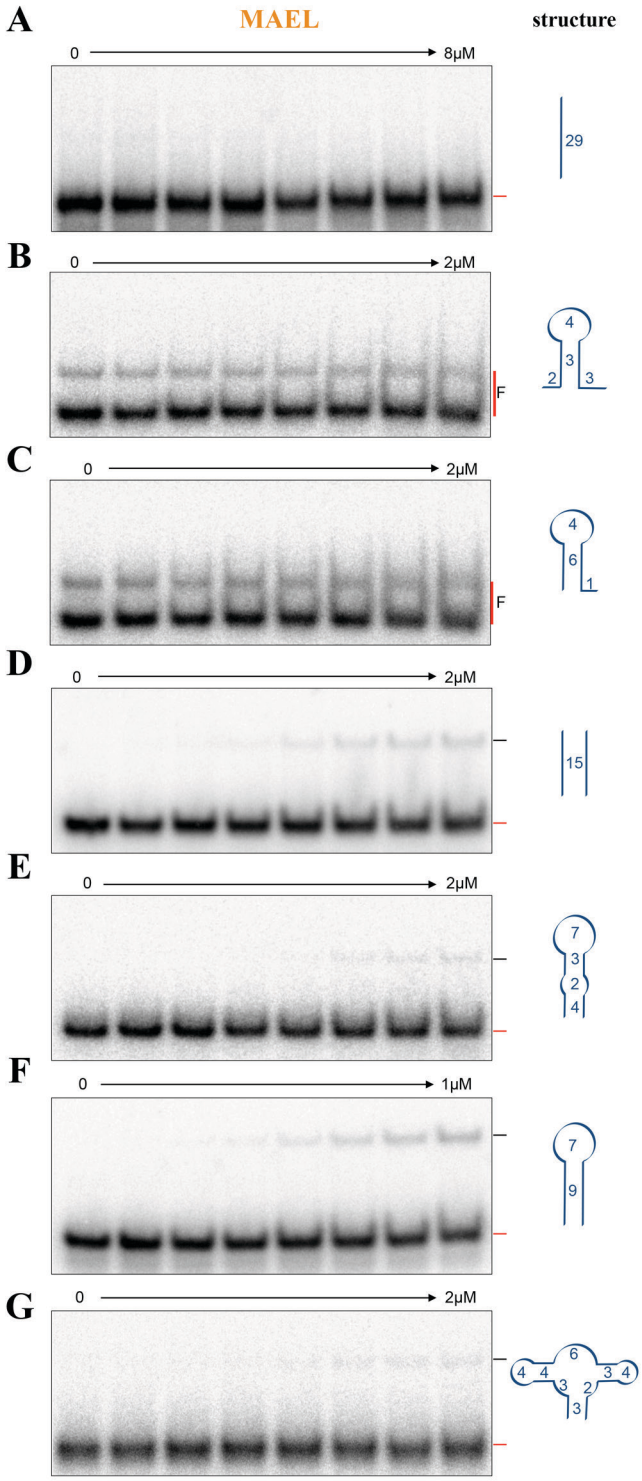
In the cell, RNA is transcribed from DNA and is often folded and associated with proteins. Its biochemical nature allows for formation of double-stranded regions, hairpins and many other secondary-structure features that are not observed for DNA [308, 309]. When double-stranded, RNA forms A-type helices, featuring a minor groove that is wide and shallow and a major groove that is narrow and deep [310]. As a result, different protein interactions are possible with dsRNA, which cannot occur with the structurally rigid dsDNA. Hairpins are one of the most commonly identified features of RNA molecules involved in protein-RNA interactions [309]. In addition to the double-stranded stem, they have single-stranded loops with unpaired bases capable of many chemical interactions with each other, other RNAs or proteins [238]. The intricacies of RNA folding capabilities are best exemplified by the ribozyme that can carry on catalytic functions by exploiting the unique biochemical nature of RNA [311].

MAEL is an important member of the piRNA pathway and specifically immunoprecipitates with piRNA precursor transcripts and transposon mRNAs [154]. Furthermore, computational approaches suggest that MAEL contains a domain with an RNAse H-like fold in the C-terminus [169, 170]. In the cell, this protein predominantly localizes to the cytoplasmic bodies believed to be involved in RNA processing [111]. In previous sections, I showed that the MAEL HMG-box does not bind dsDNA but binds strongly to the structured DNA 4WJ (Figures 16 and 18). These results, together with the known biological context (localized to RNA processing cytoplasmic bodies and associated with the piRNA pathway) suggest that the MAEL HMG-box may bind to RNA.

To test this, I designed single-stranded (ss), double-stranded (ds) and multiple-hairpin RNA substrates using *Mfold* [253] to identify the possible structural ensembles for each substrate (see methods, Table 4). While I did not observe any binding of the MAEL HMG-box to ssRNA or small hairpins (3 - 6 bp long stem with 4-nt loop) (Figure 20A, B, C), I did detect weak binding to a 15 base-pair dsRNA (Figure 20D). This substrate was longer than the double stranded regions within the stem of tested small hairpins (15 versus 3 or 6 base pairs) (Figure 20D versus B or C). I then designed two larger hairpins. The first hairpin has a 7-nt loop with a double-stranded stem disrupted by two nucleotide mismatches, and the second one was of almost identical sequence except that it had perfectly base paired stem. Because small hairpins (4-nt loop) were not bound, the loop in the new hairpins was extended to 7 bases, which provides additional unpaired bases that may contribute to interactions with protein. The stem was interrupted with two-base mismatches because one mismatch is unlikely to produce sufficient disruption to the helix. As a result, the dsRNA region next to the loop is as long as in the first small hairpin tested (Figure 20B), but the loop is longer, allowing me to evaluate its contribution. The second new hairpin lacked two base mismatches in order to determine whether the longer dsRNA region next to the loop contributes to binding. The MAEL HMG-box bound weakly to an RNA hairpin with two mismatches in the ds stem (Figure 20E) and bound better when the stem was perfectly base-paired (Figure 20F), suggesting that the larger loop and longer continuous dsRNA adjacent to it both contribute to binding. Then I have designed substrate that combined two small hairpins and ss regions, resembling three-way junction, to see whether unpaired bases in the context of two unfavorable hairpins would contribute to binding.

Figure 20: MAEL HMG-box binding to simple RNA

A) Titrations of the mouse MAEL HMG-box to the ssRNA and to B, C) RNAs that form small hairpins do not show any complex formation. Red lines next to gels highlight the free and **black** lines the shifted substrate. D) The mouse MAEL HMG-box forms complex with dsRNA of the same sequence as $dsDNA^{SRY}$ tested previously. E) The mouse MAEL HMG-box binds weakly to the hairpin with longer loop and bulge in hairpin stem. F) Binding of the mouse MAEL HMG-box to the hairpin with perfectly base-paired stem is better. G) The binding to the substrate with two unfavorable hairpins and multiple ssRNA regions is present but weaker than in (F) implying dsRNA regions are important for complex formation.


A
MAEL
structure
0
8μM
29
B
0
2μM
F
C
0
2μM
F
D
0
2μM
15
E
0
2μM
F
0
1μM
G
0
2μM

The MAEL HMG-box bound to this substrate only weakly (Figure 20G) suggesting that while ssRNA regions contribute, they require adjacent dsRNA region that is longer than 4 bases pairs. Of the tested substrates, I observed the strongest binding to the RNA hairpin with the longest continuous dsRNA stem (9 base pairs) and largest hairpin loop (7 bases) (Figure 20F). However, only ~40% of this substrate was bound at the highest protein concentration applied, prohibiting the determination of binding parameters. The observations that the MAEL HMG-box could bind dsRNA and binding got stronger in the presence of the longer hairpin loop suggested that structural features contribute to the binding. Therefore, I generated RNA 4WJ substrate using identical sequence to the tested DNA 4WJ but with RNA bases (Table 5). The MAEL HMG-box domain bound strongly to this substrate with a single complex forming at the lowest tested MAEL HMG-box concentration and multiple complexes forming at the highest concentration (Figure 21A). This was unlike the binding of HMGB1a to RNA 4WJ where progressively higher order complexes formed with the addition of protein (Figure 21C). Even though the MAEL HMG-box did not shift all the substrate, the binding strength was significantly greater compared to the single hairpin RNA substrates (Figure 21A`). To better approximate the dissociation constant, I performed competition assays using unlabeled substrates; however, the results were inconclusive (Figure 21B). Instead of the 'cold' substrate titrating away the protein from the labeled substrate, all the hot substrate shifted to large complex whose migration was reminiscent of the complex observed at highest protein concentration in previous experiment (Figure 21A).

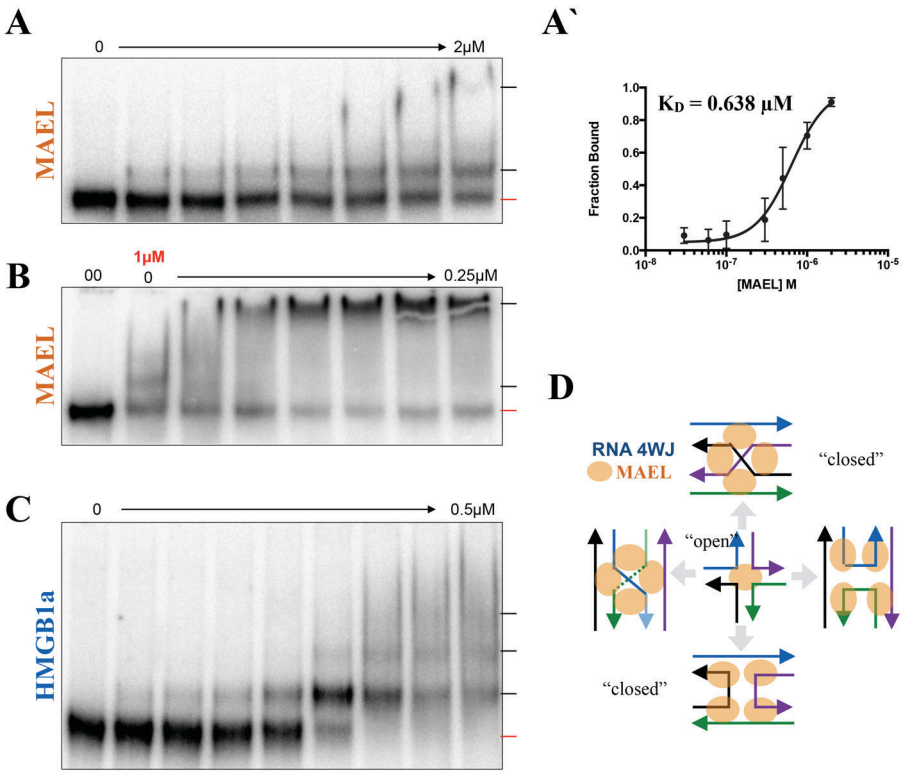


Figure 21: Binding to RNA 4WJ

A) The titration of the mouse MAEL HMG-box to RNA 4WJ showing formation of two complexes – first small and second large (substrate: free – red lines, bound – black lines). A`) Binding affinity curve shows that the binding strength is moderate (K_D = 0.638 μM). B) Addition of unlabeled RNA does not lead to competition, but unexpectedly causes increased formation of large complex and disappearance of small complex, preventing determination of competitor binding constant. The 00 lane contains only 'hot' substrate and the red value indicates amount of protein in all the following lanes. C) HMGB1a binds to RNA 4WJ strongly, progressively forming larger complexes as more protein is titrated in. D) Model describing structural ensemble of tested RNA 4WJ based on previous study [240]. Single protein is accommodated in open conformation (small complex) and multiple in closed conformation (large complex).

Section summary

Together, the above results show that the MAEL HMG-box can bind to certain RNA molecules. It does not bind ssRNA or small hairpins but forms complexes with dsRNA and larger hairpins, with preference for RNA species with ds stems featuring no mismatches in base pairing in the stem, and loops larger than 4 bases. The sequence of the dsRNA tested here was identical to the dsDNA tested previously (Figure 16D). The dsRNA, however, has different sugar that underlies its A-type helical geometry and contains uridine nucleotides. Because the type of the helix is the major difference between DNA and RNA, its helical geometry must be the reason for the observed difference of binding. Comparatively better binding to hairpins with 7 bases in the loop, as opposed to 4 bases, implies that unpaired bases in the loop also contribute to complex formation (Figure 20D vs. F). Therefore, the MAEL HMG-box may be binding near the loop region where the helical geometry is altered. This binding is further increased if the adjacent stem contains more than 6 base pair-long perfectly ds stem (Figure 21F vs. E and G).

Like its DNA counterpart, the RNA 4WJ has four helical arms and a central region where individual strands bend, deviating from the helical geometry. The same RNA 4WJ I utilized has been investigated previously with a focus on its conformational dynamics [240]. In solution, the ds arms of the junction rapidly transition from being next to each other (closed form: parallel or antiparallel orientation) to being extended outwards (open form) (Figure 21D). Instead of the cold substrate titrating away the protein from labeled substrate, all the hot substrate shifted to large complex. With additional RNA in the reaction, the ensemble may change to favor this form, leading to loss of the small

complex (Figure 21B) that resembles binding to the central region of the DNA 4WJ by a single MAEL HMG-box (Figure 21A vs. 18C). Conversely, the small complex is more abundant at low RNA substrate concentration (~1 nM), suggesting that the ensemble is favoring the open form of the substrate (Figure 21A).

The observed binding to RNA can be explained by my *in silico* analysis of the sequence and structural features of MAEL HMG-box. The arginines in the "hook" and the "propeller" are distributed such that they span approximately 270 degrees, providing sufficient rotational freedom for the rest of the domain to be accommodated in multiple ways (Figure 12). Arginine-rich peptides have been previously implicated in RNA binding [296, 312, 313]. As opposed to MAEL HMG-box, HMGB1a does not have arginine residues and, when binding to RNA 4WJ, it forms multiple complexes whose retardation corresponds to addition of more protein (Figure 21C). Taken together, the RNA binding data suggest that the MAEL HMG-box domain may be employing its arginine-rich "hook" and "propeller" regions to bind to structured RNA molecules in a complex manner, distinct from HMGB1a.

Arginines Are Crucial for MAEL HMG-box Nucleic Acid Interactions

In the previous sections, I tested the mouse MAEL HMG-box binding to several different DNA and RNA substrates. Features identified by my sequence and structural analyses, such as the arginine-rich "hook" and "propeller," may play prominent roles in both sets of interactions given their location within MAEL HMG-box. Thus, I identified residues within and outside of these regions and mutated them to alanines to evaluate their contributions to binding. Specifically, I mutated R8 and Q16 within helix-1, R23 and R25 within the "propeller" and R31 within the "hook" (Figure 22). The mutant proteins were purified using the scheme employed previously for wild-type protein (Figure 13), and any changes to their structural characteristics were identified using CD measurements (Figure 23). Of all the mutants, only the R8A secondary structure was affected. This protein was still highly helical, based on negative ellipticity in far-UV range; however, the ellipticity minimum at 222 nm has shifted, suggesting that this mutation affected the overall structure as compared to wild-type protein (Figure 23).

Mutation Q16A in the middle of helix-1 had no effect on binding to the DNA 4WJ; however, it affected the complexes formed with RNA 4WJ (Figure 24B, B`). The overall binding was decreased; small complex was lost and instead new complex at intermediate position has appeared (Figure 24B`). Mutant R8A completely abolished binding to both DNA and RNA, indicating that this residue is critically important for MAEL HMG-box's ability to interact with nucleic acid polymers (Figure 24C, C`). However, upon further inspection, the secondary structure of R8A was affected (Figure 23); therefore, it is possible that the loss of binding may instead reflect changes in protein folding.

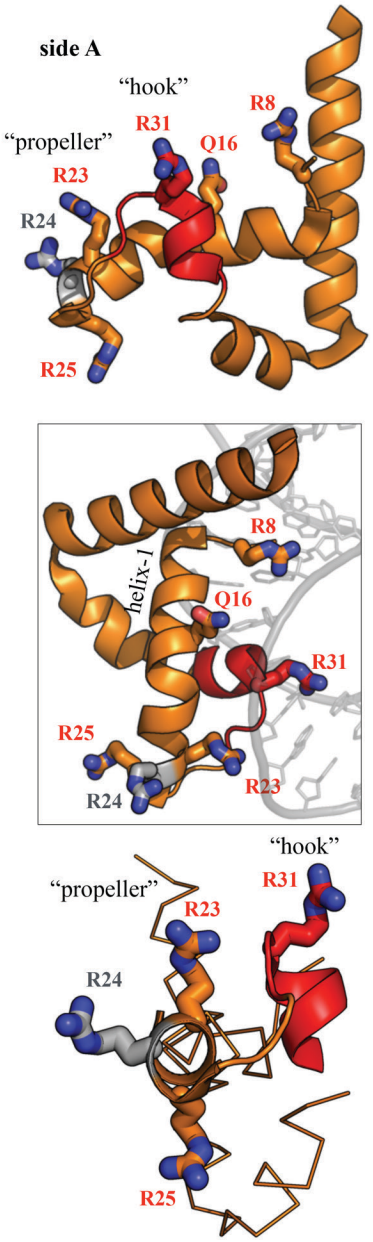


Figure 22: Mouse MAEL HMG-box mutagenesis

Multiple views highlighting the residues in important regions of the mouse MAEL HMG-box (sticks). Residues along the α-helix-1 (R8, Q16) and arginines located within the "hook" and the "propeller" regions (R23, R25, R31) were selected and mutated to alanine.

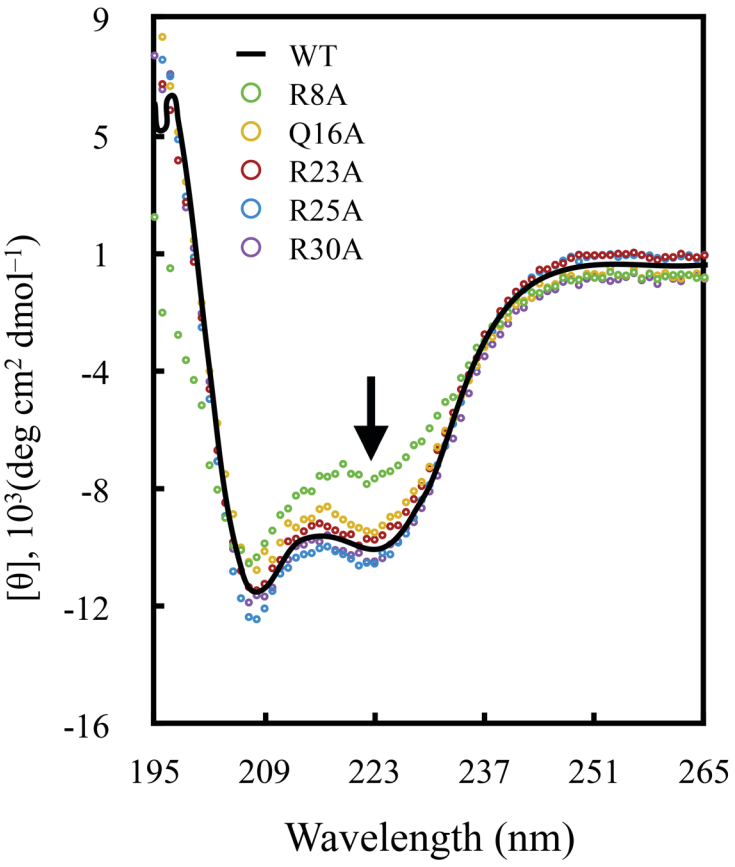
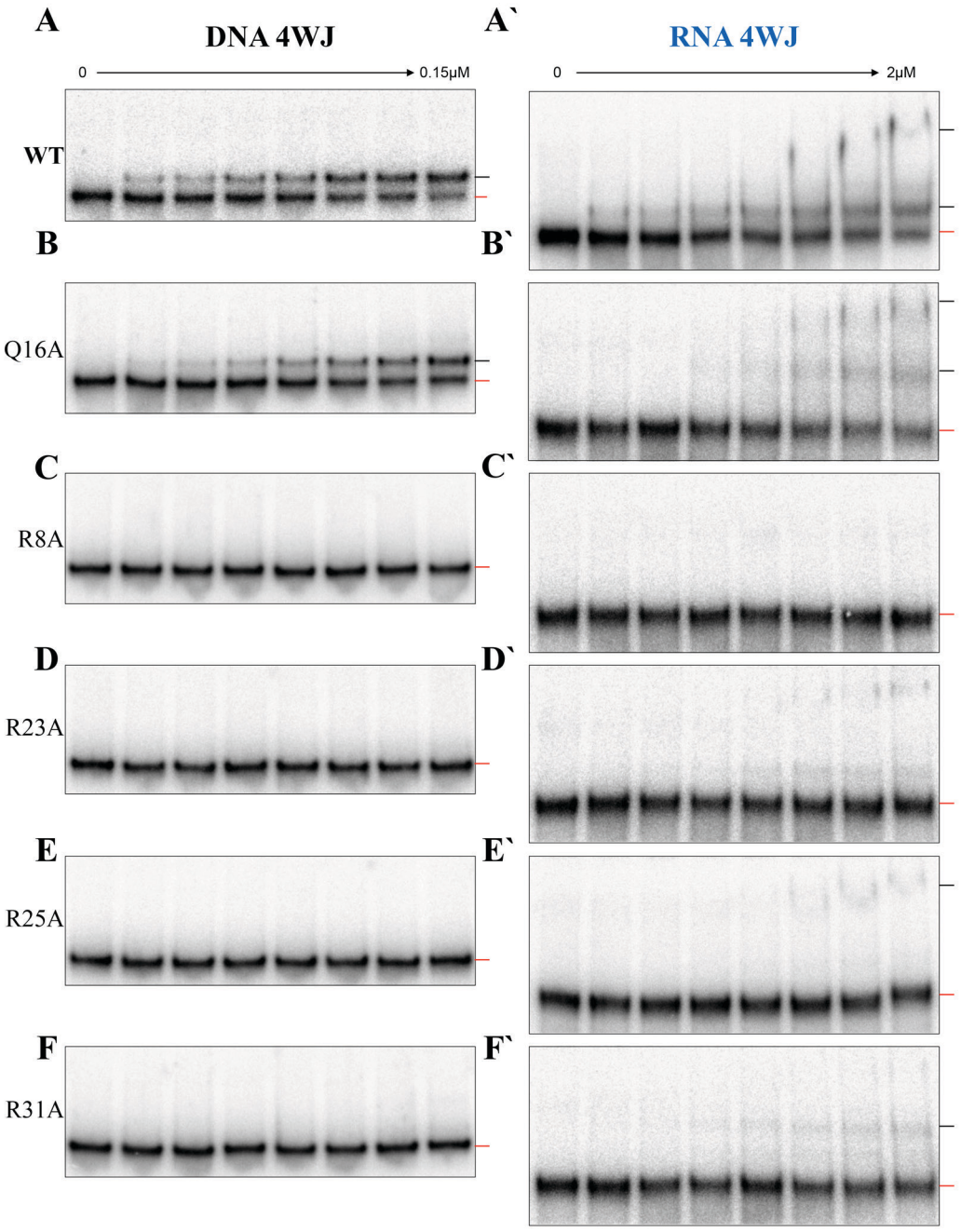


Figure 23: CD of mouse MAEL HMG-box mutants

Most of the purified mutant (colored circles) mouse MAEL HMG-box domain proteins are folded very similarly to the wild-type protein (**black** line). The folding of the R8A (**black** arrow) mutant protein is altered (change at 222 nm wavelength); however, the trace still shows negative ellipticity, indicating that protein is still α-helical.

Figure 24: Mael HMG-box mutants binding to 4WJs

A, A`) Titration of wild type mouse MAEL HMG-box to DNA and RNA 4WJ substrates. B) Mutation of glutamine (Q16A) does not affect binding to DNA 4WJ, however B`) does effect binding to RNA 4WJ. C, C`) Mutation of arginine in the N-terminus of first α-helix completely abolishes binding to both substrates. D-F) Mutation of individual arginines (R) in the "propeller" (R23A, R25A) and "hook" (R31A) regions results in the complete loss of binding to DNA 4WJ. D`-F`) Binding of these mutants to RNA 4WJ is significantly decreased compared to wild type; however, very small amounts of complexes still form. The lines represent free (red) and bound (**black**) substrate.


A
DNA 4WJ
0
0.15μM
WT
B
Q16A
C
R8A
D
R23A
E
R25A
F
R31A
A`
RNA 4WJ
0
2μM
B`
C`
D`
E`
F`

The mutants R23A, R25A and R31A completely abolished binding to DNA 4WJ (Figure 24D-F). This change is not due to an effect on protein folding (Figure 23). Binding of the three mutants (R23A, R25A, R31A) to RNA 4WJ was significantly decreased compared to wild type (Figure 24A` vs. D`-F`). Mutation R23A allowed for the formation of both the small and large complex, while only the large complex formed with the R25A mutant protein (Figure 24D`, E`). Mutation R31A abolished formation of the large complex, and instead a new complex at an intermediate position appeared (Figure 24F`).

Section summary

Mutation of an individual residue to an alanine results in replacement of the specific side chain with a methyl group that is small and cannot form H-bonds. Except for the R8A mutation, proper folding of mutant proteins was not affected (Figure 23). The R8A mutant protein is still highly helical, but its CD trace is different, suggesting that this residue plays a structural role. Plausibly, this arginine may be stabilizing the protein structure through interactions with backbone carbonyl oxygens [314], as it is located in proximity to all three helices in the unstructured N-terminal region (Figure 22). The Q16 is located in the region that in canonical HMG-boxes faces the dsDNA backbone (Figure 22). However, the MAEL HMG-box does not bind dsDNA, and binding of the Q16A mutant to the DNA 4WJ is not affected. The binding of this mutant MAEL HMG-box to RNA 4WJ, while still occurring, is reduced however (Figure 23B`), suggesting that this residue contacts the RNA substrate. The Q16 is located within the uncharged region on the side-A of the MAEL HMG-box (Figure 12C) whose uncharged nature implies that it may face the substrate instead of being exposed to the solvent. Therefore, the side chain

of Q16 may serve as an H-bond donor or acceptor and thus, be stabilizing the interaction with the RNA 4WJ [315, 316]. Arginine residues have one of the longest side chains, are positively charged, and can form H-bond networks [296]. They are also found in various RNA-binding motifs, and their binding can be modulated by methylation [317, 318]. In the MAEL HMG-box, the arginine residues are distributed within the “hook” and “propeller” motifs, and each of them individually contributes to binding to both DNA and RNA 4WJ (Figure 24). The long side chain allows for distant contact in relatively narrow spaces where the positive guanidine group can form multiple H-bonds, which has been described as the “arginine-fork” phenomenon [296]. The structural characteristics of the RNA helix, i.e. deep and narrow major grove and wide and shallow minor groove, can accommodate single as well as multiple arginine residues such as those found in the MAEL HMG-box (Figure 22). The overall mutational analysis of the MAEL HMG-box domain has revealed that arginines in the "hook" and "propeller" regions are essential for binding to the respective nucleic acid substrates, which is supported by my sequence and structure analyses and previous binding experiments that demonstrated strong complex formation with structured portions of DNA 4WJ, as well as to RNA hairpins and junctions.

MAEL HMG-box Binds Strongly to Large RNA Molecules

My in silico sequence and structure analysis combined with my extensive binding assays have laid down a solid foundation for the interaction of the MAEL HMG-box with several different substrates, including structured DNA 4WJ as well as dsRNA, RNA hairpins and RNA 4WJ. RNA 4WJs, are common structural features of cellular RNAs [319]. Furthermore, MAEL immunoprecipitates (IPs) contain large RNAs whose structures are highly complex [154]. The transcripts of transposons are often many kilo bases long, and piRNA precursor transcripts can span tens of kilo-bases [96, 179], enabling formation of nearly countless geometrical arrangements that could be bound by proteins. To explore the possibility that MAEL HMG-box directly binds to these RNAs, I interrogated MAEL IPs RNA-seq data sets for enriched RNA regions [154].

First I aligned three RNA immunoprecipitate (RIP)-Seq datasets (control: Igg, replicates: MAEL_A, MAEL_B) to the mm9 version of the mouse genome and identified enriched regions using *MACS2* software which compares read enrichment between the samples [256]. From the identified regions, I selected a subset of reads mapping to long interspersed element (LINE)-1 sequences previously shown to be enriched in the MAEL IP datasets [154]. The coordinates for each region were examined in integrated genome viewer (*IGV*) [237, 260], and narrow regions with sufficient amount of reads and peak-like appearance were selected. I have examined 114 enriched regions annotated as L1 retrotransposons sequences. Of these I selected 10 belonging to LINE1 Mus *domesticus* family 2 (L1_Md_F2), and within their coordinates, I identified 42 peak-like regions. Sequences from identified regions were then aligned (*ClustalW*) and repeatedly re-aligned (after removal of divergent sequences) until a subset with high homology to each

other was identified. As a result I have identified homologous regions on five distinct chromosomes with high sequence similarity and that are classified as L1_Md_F2 family of transposon sequences (Figure 25).

Repeated and secondary structure forming regions have been noticed within transposons [320, 321] but were never described in detail. Assuming that retrotransposon L1 RNA can be under selective pressure to retain some structural features, I attempted to determine a common secondary structure signature of identified regions. Such a common structural signature would allow for the computational search and identification of potential MAEL HMG-box target structural motifs within the tested RNA. To do this, I have employed a combination of sequence covariation and thermodynamic analysis which have been applied previously to successfully determine the structure of the yeast telomerase flexible scaffold [261]. Covariation analysis takes into account compensatory mutations between homologous sequences that are required to maintain dsRNA regions of evolutionary conserved structural elements [322], while thermodynamic analysis predicts folding based on the energetic stability [253]. Initial attempts to determine the structure of a 277-nt long piece of L1_Md_F2 with *Mfold* [253] produced a multitude of distinct thermodynamically stable structures, making it impossible to identify common regions between sequences from different chromosomes. After supplying the program with the covarying nucleotides identified with *RNAalifold* [262] software supplied with multiple sequence alignment of L1_Md_F2 regions (Figure 25A), *Mfold* produced a single structure for each region (Figure 26). It was reassuring to see that the structures originating from sequences located on different chromosomes resembled each other with only very minor variations.

Figure 25: Homology and enrichment of L1_Md_F2

A) Multiple nucleotide sequence alignment of five L1 fragments identified by searching MAEL RIP-Seq datasets. Each sequence name consists of chromosome name_DNA strand_start position. The shade of color reflects nucleotide identity going from darkest blue (completely conserved) to white (not conserved). B) The read coverage plots for each identified region. The title contains genomic coordinates and X-axis position within corresponding interval. Colors of lines correspond to different compared samples: blue - IgG, yellow - MAEL_A, green - MAEL_B.

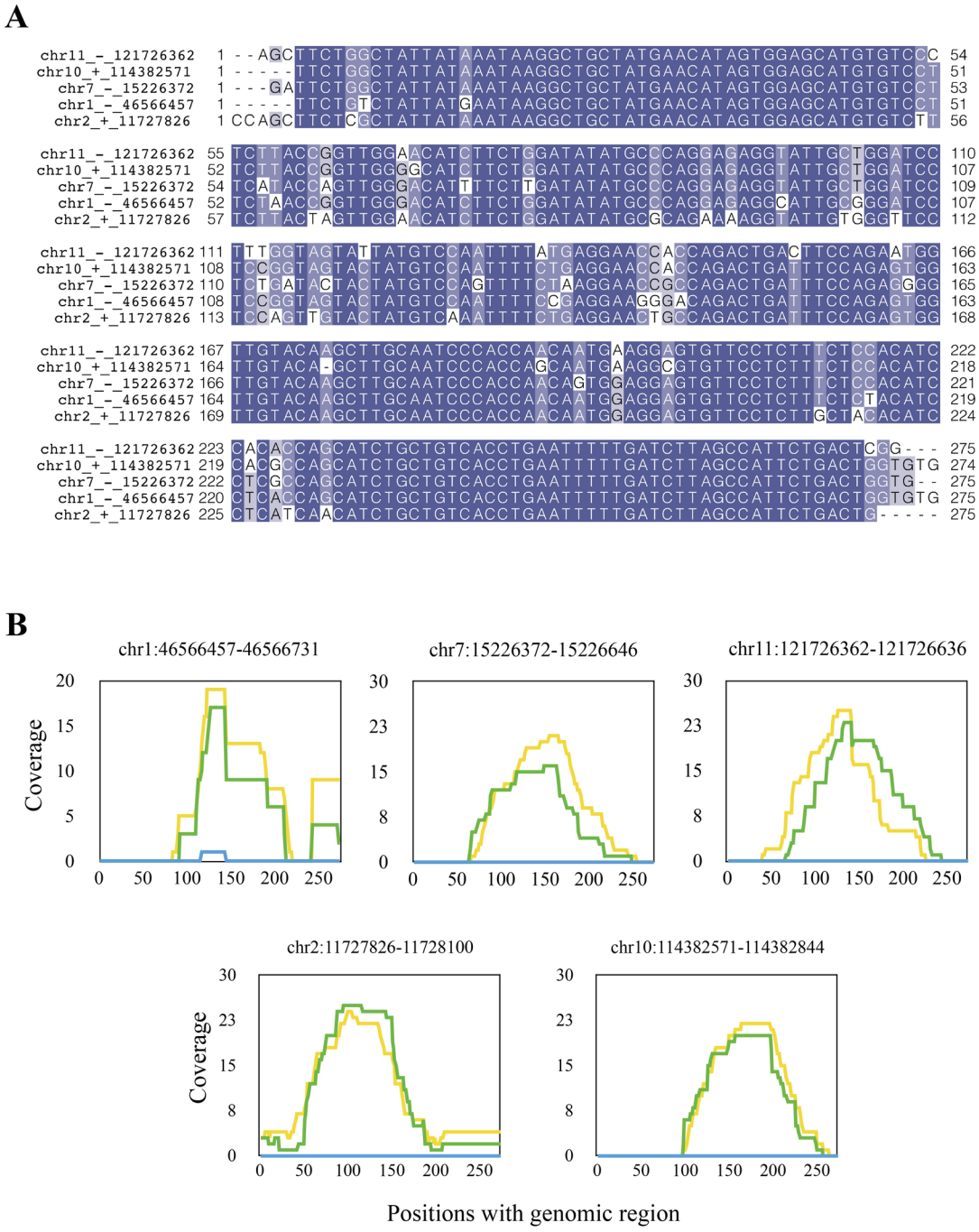
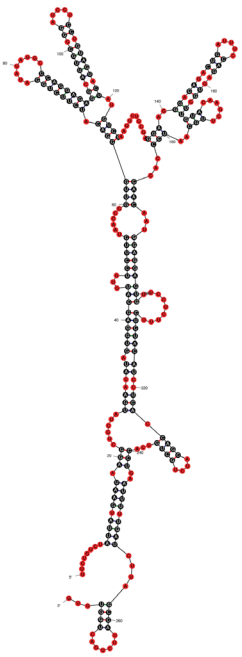

A
chr11_-_121726362
chr10_+_114382571
chr7_-_15226372
chr1_-_46566457
chr2_+_11727826
B
chr1:46566457-46566731
chr7:15226372-15226646
chr11:121726362-121726636
chr2:11727826-11728100
chr10:114382571-114382844
Coverage
Positions with genomic region

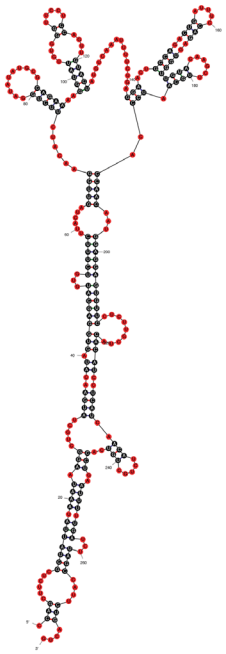
Figure 26: Predicted structures of five L1_Md_F2 regions

Secondary structures of identified regions predicted using *Mfold* supplied with constraints (see Table 6) obtained through covariation analysis using *RNAalifold* described previously [261, 262]. The coloring reflects probability of being single-stranded (red) or double-stranded (**black)** according to *Mfold*. Shown are chromosome names and *Mfold* calculated free energy of formation for each region.



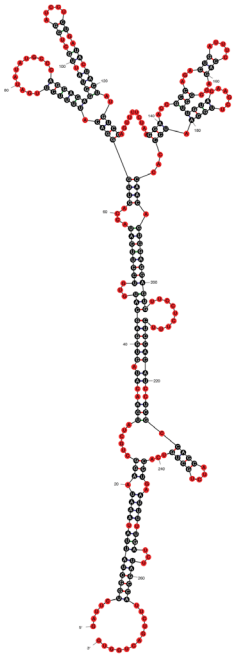
chr1

dG = -42.19



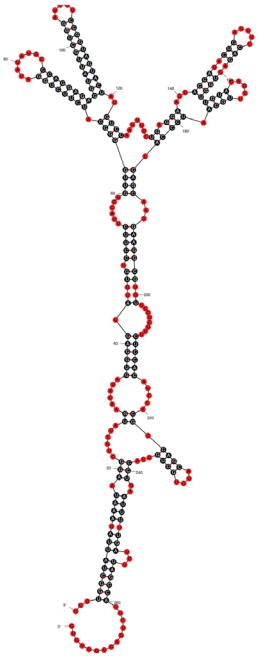
chr2

dG = -31.74



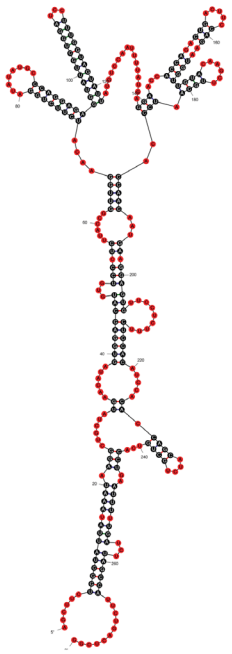
chr7

dG = -40.09



chr10

dG = -61.12



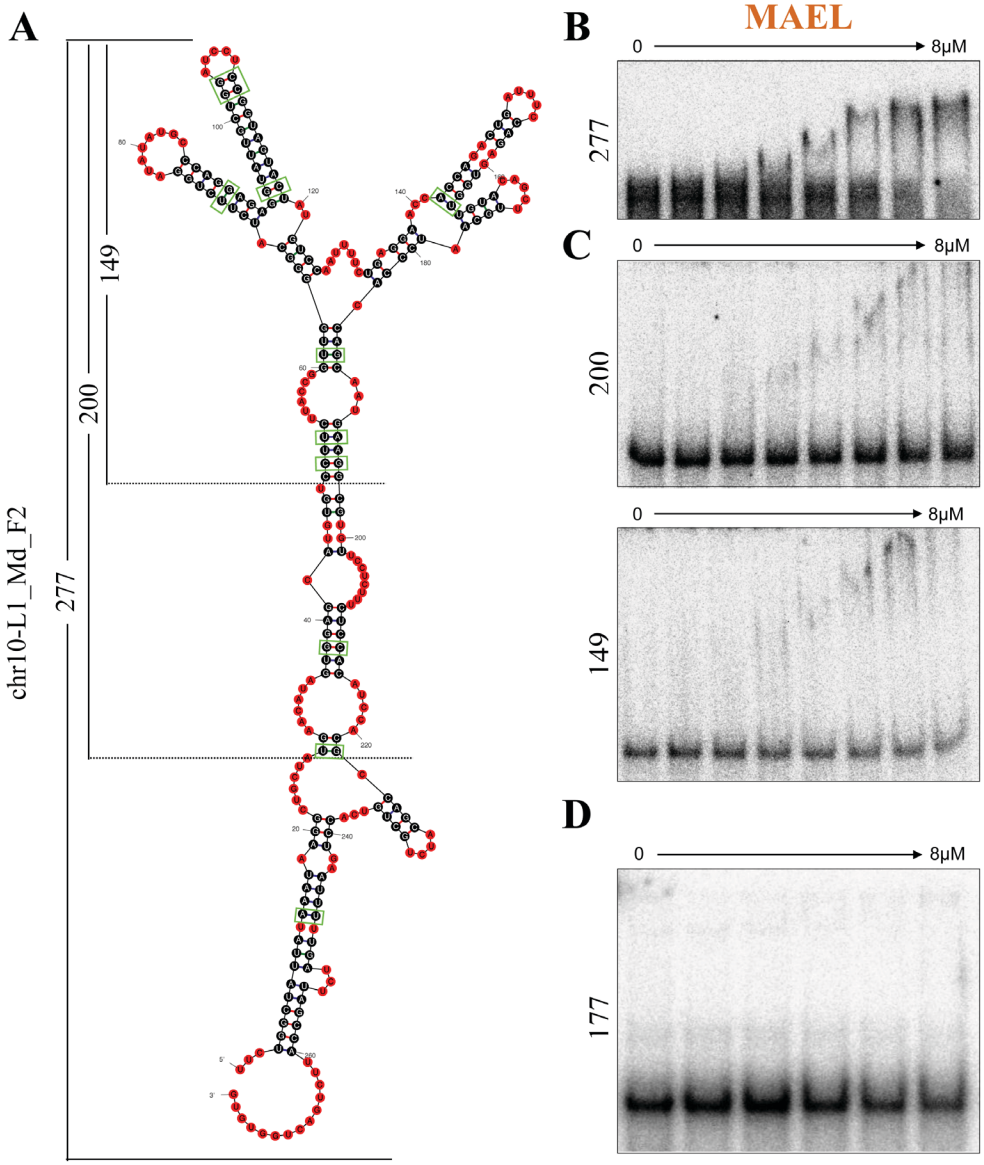
chr11

dG = -48.81

I then selected the region from chromosome 10 featuring the most energetically-stable structure and tested whether MAEL HMG-box can bind to it (Figure 27A). The MAEL HMG-box bound the full-length region (277) demonstrated by a gel shift at 0.5 μM protein (Figure 27B, lane 4 is 0.5 μM). Deletion of small segment of the double-stranded portions reduced binding, yet small amounts of complex were still observable (Figure 27C). Importantly, this interaction was specific, as RNA from genomic region not enriched in the MAEL RIP-Seq datasets was not shifted in the presence of MAEL HMG-box (Figure 27D). MAEL HMG-box domain bound the 277-nt long L1_Md_F2 RNA strongly, forming a single multi-protein-RNA complex, which became progressively more retarded in the gel shift assay with increasing MAEL HMG-box concentrations (Figure 27B). The weak complexes observed for the truncated versions (200-nt and 149-nt long) appeared to follow similar kinetics, but the majority of the substrate was unbound (Figure 27C). The binding kinetics observed for the native L1_Md_F2 substrate resemble that observed for RNA 4WJ, for which a large complex formed at higher protein and RNA concentrations (Figure 21A, B). Therefore, binding to complex RNA substrates seems to be highly cooperative, and multiple molecules are bound after passing a certain protein concentration threshold.

Figure 27: MAEL HMG-box binding to L1_Md_F2 RNA

A) Secondary structure of retrotransposon element fragment (277) from chromosome 10 and two truncations (200, 149) tested. Colors correspond to probability of being single-stranded (red) or double-stranded (**black**) according to *Mfold*. Highlighted in green squares are covarying nucleotides. B) The MAEL HMG-box binds well to 277 RNA, forming a single complex whose migration is progressively retarded with addition of protein. C). The same dynamics as in (B) is observed with shorted fragments; however, the majority of RNA has not shifted. D) The MAEL HMG-box does not bind to 177nt-long RNA that is not enriched in MAEL RIP-Seq data sets.


A
chr10-L1_Md_F2
277
200
149
B
MAEL
0
8μM
277
C
0
8μM
200
0
8μM
149
D
0
8μM
177

Section summary

The analysis of transposon RNA enriched in MAEL immunoprecipitates shows that MAEL HMG-box is capable of binding to large and specifically structured RNA molecules. This is in agreement with my previous observations, demonstrating that simple structural motifs such as RNA hairpins were bound weakly, and more complicated four-way junctions were bound strongly. While the full-length, large RNA fragment (277 nt) was bound strongly by MAEL HMG-box, the majority of 200-nt and 149-nt deletion substrates was not shifted and therefore, not bound by MAEL HMG-box. Thus, the stem that is missing in the deletion constructs likely contributes to formation of the complex. Previously, I proposed that the addition of unlabeled RNA 4WJ to the reaction causes changes in the conformational ensemble. According to this hypothesis, the double-stranded region in L1_Md_F2 may be constraining the ends of the RNA molecule, allowing for unambiguous formation of the dual hairpin region in the center (Figure 27A). In this way, the ensemble of the full-length large RNA fragment (277-nt long) structures may be smaller with greater proportion of the preferred substrate. This could explain the observed shifting of only a portion of the RNA (Figure 27B) as well as the lack of binding to short RNAs whose ensemble is composed of only unfavorable conformations (Figure 20A, B, C). Additionally, the presented structure of L1_Md_F2 region only describes secondary structural elements and no tertiary structural motifs such as pseudoknots, or interactions of secondary elements with each other. These features could provide regions analogous to those seen in RNA 4WJ that can provide deep and narrow spaces where the MAEL HMG-box can insert its arginine residues. Notably, the RNA identified from RNA-Seq data originates from retrotransposon and belongs to a

family of young and therefore likely actively transposing elements [323]. Together with its localization to cytoplasmic RNA-processing bodies associated with piRNA pathway and *in vitro* binding to structured DNA and, more importantly, RNA, this raises the possibility that the MAEL HMG-box domain contributes structure-specific RNA-binding ability to MAEL and by this way contributes to the selection of target RNAs. However, additional analysis using *in vitro* quantitative assessment of sequence and structural protein-RNA binding (RNA bind-n-Seq) [324] and *in vivo* high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) [178] approaches will be necessary to validate my findings and to identify sequence or structural RNA motifs underlying such selection.

CHAPTER IV

DISCUSSION

MAEL HMG-box as Structure-Specific RNA-binding Module

The goal of my thesis research was to gain fundamental insight into the biochemical function of MAEL, a protein indispensable for the function of the piRNA pathway. I focused on the conserved N-terminal HMG-box domain, which is crucial for the biological function of MAEL. Its classification as an HMG-box domain implied that MAEL HMG-box would bind DNA, which is the case for many canonical HMG-box domains. However, MAEL has been almost exclusively linked to the piRNA pathway; MAEL predominantly localizes to cytoplasmic bodies which are likely involved in retrotransposon-mRNA and piRNA-precursor processing (piP-bodies and chromatoid bodies) and plays a role in retrotransposon silencing, which appears to be conserved across species. Therefore, several lines of evidence suggested that the HMG-box domain of MAEL might be involved in the RNA interactions despite its HMG-box domain functional classification, which suggested MAEL HMG-box binding to DNA.

Indeed, my extensive sequence, structure and phylogenic analyses allowed me to identify unique features of the MAEL HMG-box that clearly distinguish it from the previously described SS and NSS HMG-boxes' domains. The MAEL HMG-box domain is ancient, dating back to over one billion years ago. Over this time, it has retained sequence characteristics; i.e., charged-residue distribution and the presence of multiple arginines near LINE-1, seen in vertebrates (mammals, reptiles, birds) but lost some; i.e., shortened N-terminus, in invertebrates (insects, sea urchins, fresh waster polyps). Nevertheless, structurally, the MAEL HMG-box fold did not diverge very much from canonical HMG-box domains. Comparing the mammalian MAEL HMG-boxes with canonical SS and NSS counterparts, I was able to identify arginine-rich "hook" and

“propeller” regions that are unique to MAEL HMG-box domains, indicating that the HMG-box of MAEL has acquired a new function. To identify binding substrates, I used gel-shift assays and showed that the wild-type MAEL HMG-box domain does not bind single-stranded, double-stranded, or modified dsDNA molecules even if these contain previously identified canonical HMG-box motifs. Of the DNA substrates, the MAEL HMG-box domain only bound DNA 4WJ, which is unique for its perturbed helical geometry. On the other hand, the MAEL HMG-box readily bound to RNA hairpins and formed multiple complexes with RNA 4WJ. Mutation of individual residues within the “hook” and “propeller” regions confirmed that these regions are indeed crucial for binding to the highly structured substrates – both DNA and RNA 4WJs. Because of its known role in the piRNA pathway, I have then tested MAEL HMG-box binding to an RNA fragment of the L1_Md_F2 retrotransposon identified from MAEL immunoprecipitates. The MAEL HMG-box bound strongly to this substrate but did not bind to large control RNA not enriched in the same MAEL immunoprecipitation datasets. All these observations support a model in which the MAEL HMG-box may provide its structure-specific RNA-binding abilities to the full-length protein to facilitate binding to large RNAs relevant to the piRNA pathway.

Sequence analysis shows that determinants underlying RNA-binding of the MAEL HMG-box domain (location of arginine residues within "propeller" and "hook") vary amongst mammals and insects. Structural comparison of *in silico*-predicted the fruit fly Mael HMG-box with the NMR structure of human homolog also failed to detect equivalent regions. However, because both mouse and fruit-fly domains lack the ability to bind to dsDNA; this is most likely an artifact of computational prediction algorithms.

Even though they are distantly related, both mouse and fruit-fly MAEL HMG-box domains group on the same branch of phylogenetic tree, apart from other SS and NSS HMG-box domains, implying functional conservation of their nucleic binding abilities. HMG-box domains are small domains capable of sub-specialization even with acquisition of only few mutations (i.e., SRY and SOX proteins) [220, 325], which would makes them an ideal candidate for fast adaptation to quickly accommodate ever-changing transposable elements, sequences and structures. Therefore, the sequence differences between the mouse and fruit-fly MAEL HMG-boxes could have been acquired to recognize a characteristic subset of transposon elements relevant to their host without affecting the overall conserved function of the domain. This hypothesis is supported by both different transposon repertoires of active elements [326] and their varying contributions to the intronic sequences between vertebrates and invertebrates [327]. However, experimental determination of fruit-fly HMG-box domains' tertiary structure and their biochemical interrogation is necessary to confirm this hypothesis.

Lastly, the biochemical function of the MAEL HMG-box domain *in vitro* is similar than that of HMGB1a in terms of structure-directed binding. Interestingly, in addition to their prominent structural role in the nucleus, HMGB proteins have been shown to function as a sentinel of immunogenic nucleic acids (DNA or RNA) in the innate immune response [241, 242]. A parallel presents itself where the MAEL HMG-box may have diverged to facilitate recognition of now-domesticated transposon mRNA in order to protect genome integrity during germline development. Based on this, the piRNA pathway may be considered an ancient arm of the innate immune response to protect genomes against retroviruses.

Implications of MAEL RNA-Binding for piRNA Pathway

In this thesis, I provide comprehensive set of structural and biochemical evidence (described in previous sections) describing biochemical function of the mouse MAEL HMG-box *in vitro*, all of which shine light on the enigmatic nature of MAEL protein. Presence of the structure-directed RNA-binding ability of the HMG-box agrees with MAEL localization in nuclear and cytoplasmic compartments, both of which have an abundance of RNA molecules that could be bound by the MAEL HMG-box. However, RNA immunoprecipitation experiments show that MAEL-containing complexes are preferentially enriched with large RNA molecules relevant to the piRNA pathway [154]. Some mechanistic aspects pertaining to functions of the piRNA pathway and its components have been described [328]; however, many more questions remain unanswered. A question of greatest interest to me is, "how are transcripts relevant to piRNA pathway distinguished from mRNAs?" All three, the retrotranspon RNAs, piRNA precursor RNAs, and mRNAs, are transcribed by RNA polymerase II, polyadenylated and capped [93], yet only the retrotransposon and piRNA precursor RNAs give rise to majority of piRNAs. MAEL has been observed in the vicinity of nuclear pores throughout the cytoplasm and has enriched in cytoplasmic piP-bodies believed to be centers of piRNA processing [111, 112]. MAEL localization therefore overlaps with the path RNA molecules selected for piRNA processing would take, suggesting that MAEL may be involved in shuttling retrotransposons and piRNA precursor RNA molecules. In this thesis, I showed that the *in vitro* MAEL HMG-box domain has an ability to bind large structured RNA molecules, both synthetic and native RNA found in MAEL

immunoprecipitates. Therefore, presence of an HMG-box domain with novel RNA-binding abilities in MAEL may provide the means for discrimination of RNA molecules that are relevant to piRNA pathway *in vivo*. HMG-box domains have tremendous selective capabilities and are able to distinguish nucleic acid substrates sequence-specifically (i.e., SRY, SOX proteins) or non-sequence specifically (i.e., HMGB1, Dsp1), in which case they recognize structural features. Based on my structural and biochemical analyses, the MAEL HMG-box belongs to the group that recognizes structural features of RNA.

MAEL HMG-box Also Interacts With Non-Transposon RNAs

The vastness of possible structural configurations found in RNA makes it extremely difficult to identify protein-RNA recognition motifs. Consequently, many RNA-binding proteins are considered non-specific structural binders even though they may have clear preference for a particular structure [329]. In my biochemical studies of the MAEL HMG-box, I have observed preferences for large RNA molecules originating from transposon sequences enriched in MAEL immunoprecipitates; however, I was not able to determine precise sequence or structure that would function as an ideal binding region that could be sub-cloned into a reporter construct and tested in the cell culture or other *in vivo* contexts. Nevertheless, I have identified additional large RNA molecules to which the MAEL HMG-box bound extremely well and with which it's binding followed the same dynamics as seen with L1_Md_F2 fragment (Figure 27).

One such molecule corresponds to the ribonuclease P RNA component H1 (Rpph1). The Rpph1 RNA associates with numerous proteins, and the primary function

of RNAse P holoenzyme is processing of the precursor tRNAs [330]. Initial MAEL RIP-Seq analyses by others in the lab showed high abundance of reads corresponding to Rpph1; however, later it was shown that these reads are also abundant in IgG control, suggesting that this RNA was contaminating the sequencing samples. Despite this, I have PCR-amplified and sub-cloned its cDNA into pGemT vector. Then, I digested the vector with *Spe*I enzyme and used it as run-off template in an *in vitro* transcription (IVT) reaction. The generated RNA was then folded according to same procedure as the long RNAs in previous sections (see Methods). The crystal structure of the bacterial RNAse P RNA component was solved [331, 332] and found to feature multiple RNA 4WJ [239], suggesting that MAEL HMG-box may bind it. Excitingly, the MAEL HMG-box has bound to Rpph1 RNA extremely well (Figure 28A), confirming my prediction. The unbound RNA appeared as two bands on the gel, both of which shifted as protein concentration increased, mimicking the type of interaction observed with L1_Md_F2 (Figure 27A). I have then PCR amplified the 325-nt long region, removing 67-nt plasmid sequence present in run-off transcript, and created sense and antisense RNA for the 325-nt region corresponding to only Rpph1 RNA. The final RNA was 328-nt long due to retention of 5`GGG left behind from T7 promoter sequence by T7 polymerase. The biologically relevant version of the Rpph1 RNA is encoded on the reverse-complement (RC). The shorter complement strand RNA (C) was bound equally well as the RC RNA (Figure 28B), but the overall binding of both shorter substrates was weaker than binding to the run-off RNA (Figure 28A). Just as in case of the run-off, the binding to the shortened RNAs followed same binding kinetics.

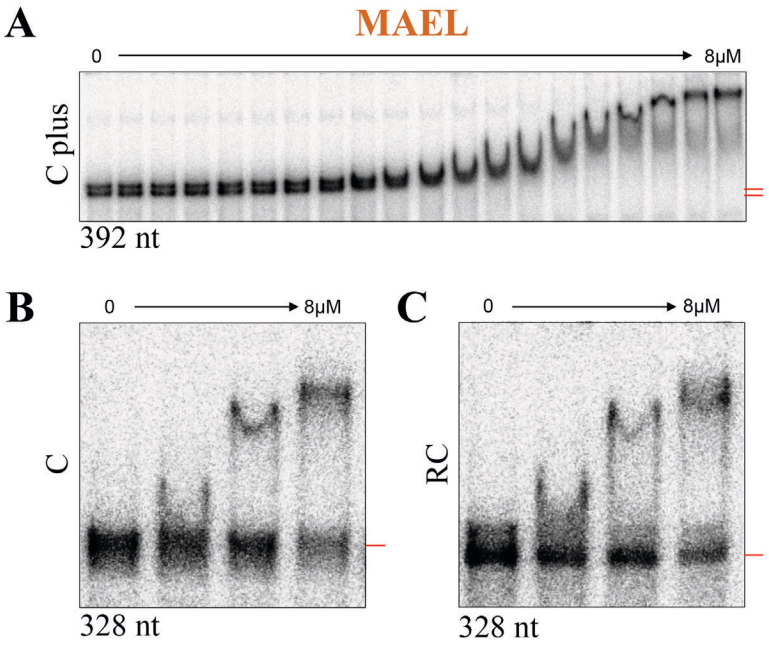


Figure 28. Binding of MAEL HMG-box to Rpph1 RNA

A) Binding to the 392nt long Rpph1 RNA. The RNA was transcribed from pGemT-Rpph1 plasmid cut with SpeI enzyme and used as run-off template in *in vitro* transcription. It contains 325-nt long Rpph1 (C - complement strand) and 67-nt long plasmid fragment (plus). B) Binding to 328-nt long Rpph1 (C) RNA transcribed from PCR product. C) Binding to 328-nt long Rpph1 (RC – reverse complement strand). The type of strand does not seem to be affecting the complex formation, suggesting that it's the nucleotide content or the structural features Rpph1 RNA forming regions recognizable by MAEL HMG-box.

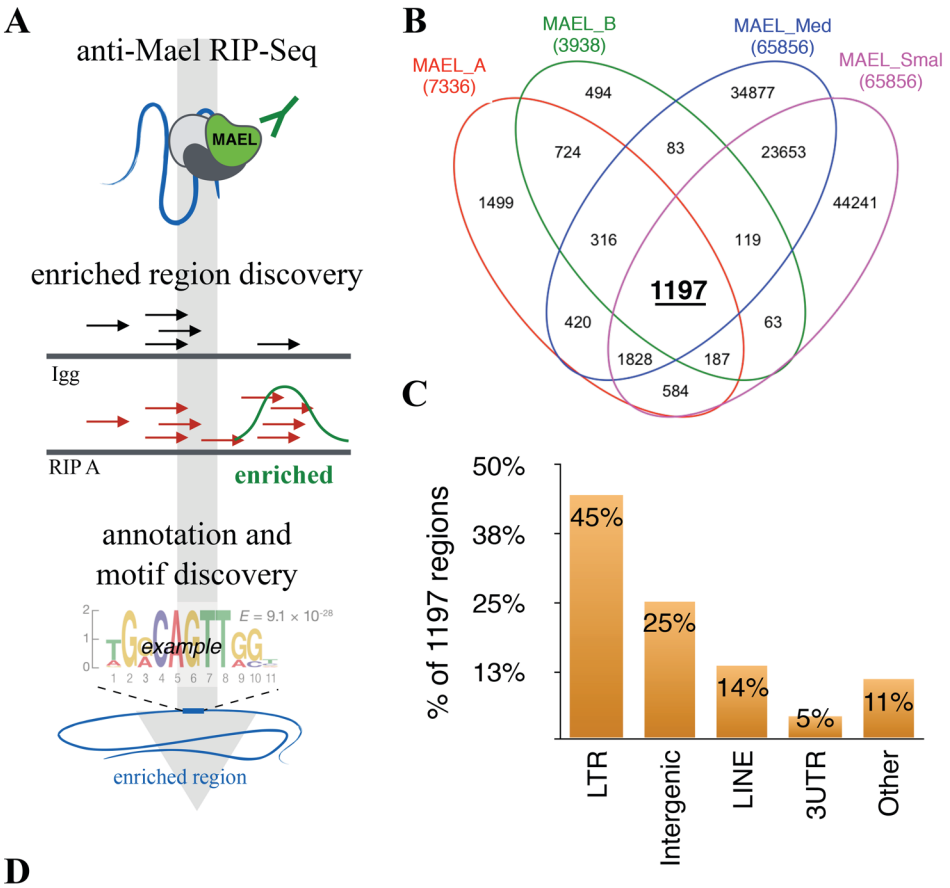
The comparatively equal binding of MAEL HMG-box to C and RC RNA implies that it may be recognizing structural features that are common between the two substrates rather than specific sequences. Just as in the case of L1_Md_F2 and its truncations, the relatively better binding to run-off RNA of Rpph1 may be occurring as a consequence of improved folding in presence of the 67-nt long plasmid sequence. At this point I cannot state whether binding to Rpph1 RNA by MAEL HMG-box is of relevance to the overall MAEL function in the piRNA pathway because it is not enriched in the MAEL RIP-Seq datasets. However, because this RNA was bound well, and unlike L1_Md_F2, tertiary structure of its homologue has been experimentally determined [331, 332]. In the future it may prove a useful tool for identification of specific structural features that are bound by the MAEL HMG-box.

Searching For Sequence Signatures in MAEL RIP-Seq Data

In my thesis, I have described MAEL HMG-box binding to the L1_Md_F2 retrotransposon region that was enriched in MAEL RIP-Seq Datasets. My analysis was limited to three published RNA-seq data sets (Igg, MAEL_A, MAEL_B), and I have analyzed RNA of only a single type of enriched sequences. Nevertheless, MAEL HMG-box was able to distinguish enriched L1_Md_F2 RNA from RNA that was not enriched in these datasets (Figure 27). This suggested that even though MAEL RIP contained many other proteins [154], MAEL might be the one helping in selection of the enriched RNA molecules.

Figure 29: Discovering sequence signatures in the MAEL RIP-Seq Data

A) The schematic representation the major steps taken in identification of sequence signatures within MAEL RIP-Seq datasets. B) Venn diagram generated by DiffBind R program summarizing number of discovered regions within and between each sample. There were total of 1197 regions common to all data sets. C) The percentile distribution of identified regions showing the categories with most members. D) The sequence motif logos discovered (MEME) within annotated groups with >5% of total sequences. Shown is total number of regions within each group, the motif logo reflecting enrichment of each nucleotide at given position within motif, and associate E-value reflecting probability of being present by chance.



D

regions	motif	E-value
1197 all		1.4e-583
534 LTR		4.6e-357
304 Intergenic		2.7e-147
163 LINE		3.2e-335
54 3`UTR		1.7e-041

Therefore, I have decided to query the same and additional MAEL RIP-Seq datasets (MAEL_Med, MAEL_Smal) using genome-wide computational approaches with a goal of identifying conserved sequence signature within enriched regions that may be underlying the structural determinants that the MAEL HMG-box recognizes (Figure 29A). I have used *MACS2* [256] to identify enriched regions, and *DiffBind R* package [333, 334] to identify 1197 regions common amongst all datasets (Figure 29B).

The identified regions were then annotated using *annotatePeaks* program from *Homer* suite [258]. The majority of the regions were annotated as retrotransposons (long terminal repeat (LTR) = 45%, LINE = 14%), some as unknown intergenic sequences (25%), and the rest corresponded to genic features such as 3`UTR (5%) and others (11% includes 5`UTR, introns, exons, transcription start sites, transcription termination sites, pseudogenes). (Figure 29C). To determine whether any sequence motifs are present within identified regions, I have employed the *MEME* program [335] from the *MEME* Suite [336]. I used a setting that searched for one occurrence of up to 30-nt long motif per each sequence within all as well as annotated subset with > 5% of discovered regions. The *MEME* program identified 18-nt and 29-nt long motifs, all of which were overrepresented with adenine (A) and guanine (G) bases in vast majority of positions (Figure 29D). This was the case for the 1197 genomic regions as well as for each annotation category; however, the bit values (motif Y-axis) that indicate prevalence of nucleotide at the given position within set of tested regions never reached its maximum (value of 2), suggesting low conservation of these sequences. The nucleotides within identified motifs also periodically varied, switching between A and G. Presence of periodic G nucleotides is reminiscent of sequences that give rise G-quadruplex structures

observed in telomeric DNA regions [337]. More importantly G-quadruplexes were recently identified in nascent RNA transcripts [338], 5` untranslated regions [339], and transposable elements [340, 341]. In all these contexts they are thought to serve important regulatory functions by facilitating formation of specific tertiary structures that are either bound by proteins or contribute to RNA tertiary structures [238, 342]. This indicated G-quadruplex forming seqeunce may also be important for MAEL HMG-box binding; therefore, I used *quadruplex forming G-rich sequences* (*QGRS*) software [343] to evaluate my large RNA substrates for their presence. Excitingly, the three of five aligned L1_Md_F2 regions had single predicted G-quadruplex, and sequence from chr10 that was tested for binding had two. Furthermore, compliment strand (C) of Rpph1 RNA had four, and reverse-compliment had six sequences that were predicted to form Q-quadruplex structures. These observations suggested that the MAEL HMG-box might be binding to large RNAs, recognizing the G-quadruplex structures. Therefore, I have tested the regions from each substrate predicted to form these structures and tested them in gel shift assays. Unfortunately, I was not able to detect any binding to short DNA or RNA molecules corresponding to G-quadruplex. My previous tests with truncated L1_Md_F2 substrates and with RNA 4WJ implied that additional RNA in binding reaction contribute to MAEL HMG-box binding. Perhaps G-quadruplex sequences need to be in the context of larger RNA molecule to be recognized by the MAEL HMG-box. In such context they may give rise to other RNA geometries that can better accommodate the arginine residues essential for MAEL HMG-box binding.

While my computational analysis of MAEL RIP-Seq datasets pointed me in the direction of G-quadruplex sequences, I cannot exclude the possibility that other sequence

or structural RNA features may exit for which MAEL HMG-box has preference. In order to address this possibility, HITS-CLIP [178] experiments utilizing UV-crosslinking to identify RNA directly bound to MAEL protein are necessary. These could be combined with recently described RNA bind-N-seq technique, which combines large-scale sequencing with RNA-binding assays and structural predictions to identify features bound by RNA-binding proteins [324]. Lastly, the importance of identified novel features of the MAEL HMG-box (arginine rich "hook" and "propeller") would be best demonstrated *in vivo* through mouse transgenesis. To that end, I have generated construct deleting the MAEL HMG-box as well as those with R31A point mutations to be used in doxycycline inducible over-expression paradigmin the future [344].

TABLES

Table 1 is located on p. 23 to preserve continuity of the section

Table2. Cloning oligonucleotides

Length	Type	Tm	Sequence 5`-->3`	Gene/Fragment	Notes
30	Fw	60	AAATA**GGATCC**ATAGAGGGCCATGTCAAGC	Mm SRY HMG	To amplify Mm SRY HMG without start codon. Includes BamH1 site for subcloning into pGex6P-2
37	Rew	60	AAATA**GCGGCCGC**ACTCCTCTGTGACACTTTAGCCCT	Mm SRY HMG	To amplify Mm SRY HMG endign at residue 86. Iincludes NotI site for subcloning into pGext6P-2.
31	Fw	60	AAATA**GGATCC**ATAGCTCCTAAGAAGCATAG	Dm Mael HMG	To amplify Dm MAEL HMG without start codon. Includes BamH1 site for subcloning into pGex6P-2.
30	Rew	60	AAATA**GCGGCCGC**ATCAAGCTCTCGGTGGC	Dm Mael HMG	To ampligy Dm MAEL HMG ending at residue 86. Includes NotI sire for subcloning into pGex6P-2.
23	Fw	60	AAATA**GGATCC**GTGCCCAACCGC	Mm Mael HMG	To amplify Mm MAEL HMG withou start codon. Includes BamH1 site for subcloning into pGex6P-2.
28	Rew	60	AAATA**GCGGCCGC**TGGCCTCCTCAGTGG	Mm Mael HMG	To amplify Mm MAEL HMG ending at residue 86. Includes Not1 site for subcloning into pGex6P-2.
35	Fw	59	AAATA**GGATCC**ATAGGCAAAGGAGATCCTAAAAAG	Mm HMGB1 Box A	To amplify Mm HMGB1 Box A without start codon. Includes BamH1 site for cloning into pGex6P-2.
30	Rew	59	AAATA**GCGGCCGC**TTTGGTCTCCCCTTTGG	Mm HMGB1a Box A	To amplify Mm HMGB1 Box A ending at residue 86. Includes Not1 site for cloning into pGex6P-2.

NNNN - restriction enzyme recognition site
NNNN - additional sequence for restriction enzyme
Fw - forward
Rew -reverse

Table 3. DNA substrate sequences

Length (nt)	Type	Sequence 5`-->3`	Substrate
117	Fw	AAATATAATACGACTCACTATAGGGCTACCTACTGCAGCCGGTTACCTTTTATATAAGACAGTC GTAAACAAAGTAGGGGTGATTGTTCAGCCCTATATGTGAGTCGTATTATATTT	dsDNA
	Rew	AAATATAATACGACTCACATATAGGGCTGAACAATCACCCCTACTTTGTTTACGACTGTCTTATA TAAAAGGTAACCGGCTGCAGTAGGTAGCCCTATAGTGAGTCGTATTATATTT	
48	Fw	AACAAAGTAGGGGTGATTGTTCAGAACAAAGTAGGGGTGATTGTTCAG	dsDNA
	Rew	CTGAACAATCACCCCTACTTTGTTCTGAACAATCACCCCTACTTTGTT	
48	Fw	ATGAGGAACTTACCAGTTTTTCTTATCAATGAAACAATAACAAAGCGC	dsDNA
	Rew	GCGCTTTGTTATTGTTTCATTGATAAGAAAAACTGGTAAGTTCCTCAT	
40	Fw	GTGTAGTCGTGATAGGAAAAAAATGCAGGGGGTTATAGGG	dsDNA
	Rew	CCCTATAACCCCCTGCATTTTTTTCCTATCACGACTACAC	
26	Fw	AAAGGGTTAGGGTTAGGGTTAGGGAA	ssDNA
24	Fw	AACAAAGTAGGGGTGATTGTTCAG	dsDNA
	Rew	CTGAACAATCACCCCTACTTTGTT	
16	Fw	CAGTCGACGTCGTGAC	dsDNA
	Rew	GTCACGACGTCGACTG	
16	Fw	CAGTCGA[5me-dC]GTCGTGAC	methylated dsDNA
	Rew	GTCACGA[5me-dC]GTCGACTG	
15	Fw	GCGCAACAATGCCGG	dsDNA
	Rew	CCGGCATTGTTGCGC	
38	A	CCGACAGGACTGTCAACCAGGTAATATACCACTTGCGG	DNA 4WJ
35	B	CCGCAAGTGGTATATTACCTGGTACGCGTTCACGG	
27	C	CCGTGAACGCGTGGTGCGAATCGG	
24	D	CCGATTCGCACCTGACAGTCCTGTCGG	

Fw - forward
Rew - reverse
A-D - strands of 4WJ

Table4. RNA substrate sequences

Length	Type	Sequence 5` --> 3`	Substrate	dG (kcal/mol)*	No. structures*
28	ss	GGGAAAGGGUUAGGGUUAGGGUUAGGGAA	ssRNA	> 1.5	0
15	Fw	GCGCAACAAUGCCGG	dsRNA	n.a.	n.a.
15	Rew	CCGGCAUUGUUGCGC			
15	ss	GCGCAACAAUGCCGG	hairpin 1	-1.3	3
17	ss	GGGACCAGAAGGUCCCG	hairpin 2	-12.1	1
25	ss	GGAUCUCUGCACACAACAGUUGUCC	hairpin 3	-5.7	1
25	ss	GGAUCUCUGCACACAACAGagGUCC	hairpin 4	-12.5	1
39	ss	CCCAACACCCGCAAGGGUCCACACGGGACUCCCCACGGG	hairpin 5	-16.8	1
38	A	[dC]CGACAGGACUGUCAACCAGGUAAUAUACCACUUGCG[dG]	RNA 4WJ	n.a.	n.a.
35	B	[dC]CGCAAGUGGUAUAUUACCUGGUACGCGUUCACG[dG]			
27	C	[dC]CGUGAACGCGUGGUGCGAAUCG[dG]			
24	D	[dC]CGAUUCGCACCUGACAGUCCUGUCG[dG]			

[dN] - deoxy nucleotide (DNA)
* based on Mfold analysis
ss - single stranded
Fw - forward
Rew - rewerse
A-D - strands of 4WJ

Table5. PCR and IVT transription oligonucleotides

Length	Type	Tm	Sequence 5`-->3`	Gene/Fragment	Notes
25	Fw	66	GTTCTGTCTTCCTGTATAGGCTGGC	chr10_L1_Md_F2-1513	To amplify unique genomic fragment containing region of L1_Md_F2 on chromosome 10.
26	Rew	66	GGTCCAAATGCCAACACAGTCTTAAG	chr10_L1_Md_F2-1513	To amplify unique genomic fragment containing region of L1_Md_F2 on chromosome 10.
25	Fw	62	TTCTGGCTATTATAAATAAGGCTGC	chr10_L1_Md_F2-274	To amplify only the peak of L1_Md_F2 on chromosome 10 - referred to as 277.
19	Rew	62	CACACCAGTCAGAATGGCT	chr10_L1_Md_F2-274	To amplify only the peak of L1_Md_F2 on chromosome 10 - referred to as 277.
21	Fw	62	TGAACATAGTGGAGCATGTGT	chr10_L1_Md_F2-197	To amplify truncation of L1_Md_F2 region on chromosome 10 -referred to as 200.
17	Rew	62	GGCGTGGATGTGGAGAA	chr10_L1_Md_F2-197	To amplify truncation of L1_Md_F2 region on chromosome 10 -referred to as 200.
18	Fw	62	CCTTCTTACCGGTTGGGG	chr10_L1_Md_F2-146	To amplify truncation of L1_Md_F2 region on chromosome 10 -referred to as 147.
19	Rew	61	TTCATTGCTGGTGGGATTG	chr10_L1_Md_F2-146	To amplify truncation of L1_Md_F2 region on chromosome 10 -referred to as 147.
45	Fw	62	TAATACGACTCACTATAGGGTTCTGGCTATTATAAATAAGGCTGC	T7_chr10_L1_Md_F2-274	To in vitro transcribe 274 region of L1_Md_F2. The resulting product is 277 nucleotides due to GGG addition by T7.
41	Fw	62	TAATACGACTCACTATAGGGTGAACATAGTGGAGCATGTGT	T7_chr10_L1_Md_F2-197	To in vitro transcribe 197 region of L1_Md_F2. The resulting product is 200 nucleotides due to GGG addition by T7.
38	Fw	62	TAATACGACTCACTATAGGGCCTTCTTACCGGTTGGGG	T7_chr10_L1_Md_F2-146	To in vitro transcribe 146 region of L1_Md_F2. The resulting product is 149 nucleotides due to GGG addition by T7.
33	Fw	65	GGGGATGAAAAATTCATCTAATTTAGTACAATG	chr2_noPeak-174	To amplify control genomic region not found in MAEL RIP - referred to as 177.
23	Rew	65	GTTTGGGCAGGAGCATACATACC	chr2_noPeak-174	To amplify control genomic region not found in MAEL RIP - referred to as 177.
49	Fw	62	TAATACGACTCACTATAGGGGGGGATGAAAAATTCATCTAATTTAGTAC	T7_chr2_noPeak-174	To in vitro transcribe 174 region not found in MAEL RIP. The resulting product is 177 nucleotides due to GGG addition by T7.
46	Fw	n/a	TAATACGACTCACTATAGGGAAAGGGTTAGGGTTAGGGTTAGGGAA	29nt-ssRNA	To in vitro transcribe synthetic RNA. Watson.
46	Rew	n/a	TTCCCTAACCCTAACCCTAACCCTTTCCCTATAGTGAGTCGTATTA	29nt-ssRNA	To in vitro transcribe synthetic RNA. Crick.

TAATACGACTCACTATAGGG - T7 promoter sequence
ss - ssRNA
Fw - forward
Rew - reverse

Table6. Long RNA sequences

Length	sequence 5` --> 3`	chr	start	end	Name
277	GGGUUCUGGCUAUUAUAAAUAAGGCUGCUAUGAACAUAGUGGAGCAUGUGU CCUUCUUACCGGUUGGGGCAUCUUCUGGAUAUAUGCCCAGGAGAGGUAUUG CUGGAUCCUCCGGUAGUACUAUGUCCAAUUUUCUGAGGAACCACCAGACUGA UUUCCAGAGUGGUUGUACAGCUUGCAAUCCCACCAGCAAUGAAGGCGUGUU CCUCUUUCUCCACAUCCACGCCAGCAUCUGCUGUCACCUGAAUUUUUGAUCU UAGCCAUUCUGACUGGUGUG	chr10	114382571	114382844	L1_Md_F2
200	GGGUGAACAUAGUGGAGCAUGUGUCCUUCUUACCGGUUGGGGCAUCUUCUG GAUAUAUGCCCAGGAGAGGUAUUGCUGGAUCCUCCGGUAGUACUAUGUCCA AUUUUCUGAGGAACCACCAGACUGAUUUCCAGAGUGGUUGUACAGCUUGCAA UCCCACCAGCAAUGAAGGCGUGUUCCUCUUUCUCCACAUCCACGCC	chr10	11438598	114382794	L1_Md_F2
149	GGGCCUUCUUACCGGUUGGGGCAUCUUCUGGAUAUAUGCCCAGGAGAGGUA UUGCUGGAUCCUCCGGUAGUACUAUGUCCAAUUUUCUGAGGAACCACCAGAC UGAUUUCCAGAGUGGUUGUACAGCUUGCAAUCCCACCAGCAAUGAA	chr10	114382619	114382764	L1_Md_F2
177	GGGGGGGAUGAAAAAUUCAUCUAAUUUAGUACAAUGUAAAUAUAAAGAGUUU GGAAAUUUUUGAAAACUUGAGGUCAUGUUAAAAAAGCAACAACUCUCCUUUU AGAUUGUCCUAUUUGUGUCUUGAAAUUCCUAUGCUGUUCUAAACCCAUUUGG UAUGUAUGCUCCUGCCCAAAC	chr2	11847611	11847784	none

GGG -part of the T7 promoter not included in genomic region

Table7. Mfold constrains.

chr1	chr2	chr7	chr10	chr11
F 102 110 1	F 107 115 1	F 104 112 1	F 102 110 1	F 105 113 1
F 103 109 1	F 108 114 1	F 105 111 1	F 103 109 1	F 106 112 1
F 94 118 1	F 99 123 1	F 96 120 1	F 94 118 1	F 97 121 1
F 72 89 1	F 77 94 1	F 74 91 1	F 72 89 1	F 75 92 1
F 14 248 1	F 19 253 1	F 16 250 1	F 14 247 1	F 17 251 1
F 50 196 1	F 55 201 1	F 52 198 1	F 50 195 1	F 53 199 1
F 52 194 1	F 57 199 1	F 54 196 1	F 52 193 1	F 55 197 1
F 142 164 1	F 147 169 1	F 144 166 1	F 142 164 1	F 145 167 1
F 38 214 1	F 43 219 1	F 40 216 1	F 38 213 1	F 41 217 1
F 28 223 1	F 33 228 1	F 30 225 1	F 28 222 1	F 31 226 1
F 61 188 1	F 66 193 1	F 63 190 1	F 61 187 1	F 64 191 1

BIBLIOGRAPHY

[1] Fujiwara T, Dunn NR and Hogan BL (2001) Bone morphogenetic protein 4 in the extraembryonic mesoderm is required for allantois development and the localization and survival of primordial germ cells in the mouse. Proceedings of the National Academy of Sciences of the United States of America 98: 13739-13744.

[2] Ying Y and Zhao GQ (2001) Cooperation of endoderm-derived BMP2 and extraembryonic ectoderm-derived BMP4 in primordial germ cell generation in the mouse. Developmental biology 232: 484-492.

[3] Ying Y, Liu XM, Marble A, Lawson KA and Zhao GQ (2000) Requirement of Bmp8b for the generation of primordial germ cells in the mouse. Molecular endocrinology 14: 1053-1063.

[4] Lawson KA, Dunn NR, Roelen BA, Zeinstra LM, Davis AM, et al. (1999) Bmp4 is required for the generation of primordial germ cells in the mouse embryo. Genes & development 13: 424-436.

[5] Seki Y, Yamaji M, Yabuta Y, Sano M, Shigeta M, et al. (2007) Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. Development 134: 2627-2638.

[6] Seki Y, Hayashi K, Itoh K, Mizugaki M, Saitou M, et al. (2005) Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. Developmental biology 278: 440-458.

[7] Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, et al. (2002) Epigenetic reprogramming in mouse primordial germ cells. Mechanisms of development 117: 15-23.

[8] Surani MA, Durcova-Hills G, Hajkova P, Hayashi K and Tee WW (2008) Germ line, stem cells, and epigenetic reprogramming. Cold Spring Harbor symposia on quantitative biology 73: 9-15.

[9] Extavour CG and Akam M (2003) Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. Development 130: 5869-5884.

[10] Chaillet JR, Vogt TF, Beier DR and Leder P (1991) Parental-specific methylation of an imprinted transgene is established during gametogenesis and progressively changes during embryogenesis. Cell 66: 77-83.

[11] Sapienza C, Peterson AC, Rossant J and Balling R (1987) Degree of methylation of transgenes is dependent on gamete of origin. Nature 328: 251-254.

[12] Reik W, Collick A, Norris ML, Barton SC and Surani MA (1987) Genomic imprinting determines methylation of parental alleles in transgenic mice. Nature 328: 248-251.

[13] Monk M, Boubelik M and Lehnert S (1987) Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development 99: 371-382.

[14] Kafri T, Ariel M, Brandeis M, Shemer R, Urven L, et al. (1992) Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes & development 6: 705-714.

[15] McLaren A (1984) Meiosis and differentiation of mouse germ cells. Symposia of the Society for Experimental Biology 38: 7-23.

[16] Obata Y, Kaneko-Ishino T, Koide T, Takai Y, Ueda T, et al. (1998) Disruption of primary imprinting during oocyte growth leads to the modified expression of imprinted genes during embryogenesis. Development 125: 1553-1560.

[17] Bennett MD (1977) The time and duration of meiosis. Philosophical transactions of the Royal Society of London Series B, Biological sciences 277: 201-226.

[18] Hata K, Okano M, Lei H and Li E (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development 129: 1983-1993.

[19] Bourc'his D and Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96-99.

[20] Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, et al. (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature 429: 900-903.

[21] De La Fuente R, Baumann C, Fan T, Schmidtmann A, Dobrinski I, et al. (2006) Lsh is required for meiotic chromosome synapsis and retrotransposon silencing in female germ cells. Nature cell biology 8: 1448-1454.

[22] Orgel LE and Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284: 604-607.

[23] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

[24] Doolittle WF and Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601-603.

[25] Bestor TH (2003) Cytosine methylation mediates sexual conflict. Trends in genetics : TIG 19: 185-190.

[26] Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, et al. (2008) DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes & development 22: 908-917.

[27] Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, et al. (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Molecular cell 31: 785-799.

[28] Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

[29] Mc CB (1950) The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences of the United States of America 36: 344-355.

[30] Feschotte C and Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annual review of genetics 41: 331-368.

[31] Eickbush TH and Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus research 134: 221-234.

[32] Havecker ER, Gao X and Voytas DF (2004) The diversity of LTR retrotransposons. Genome biology 5: 225.

[33] McCarthy EM and McDonald JF (2004) Long terminal repeat retrotransposons of Mus musculus. Genome biology 5: R14.

[34] Babushok DV and Kazazian HH, Jr. (2007) Progress in understanding the biology of the human mutagen LINE-1. Human mutation 28: 527-539.

[35] Belancio VP, Deininger PL and Roy-Engel AM (2009) LINE dancing in the human genome: transposable elements and disease. Genome medicine 1: 97.

[36] Wagner A (2009) Transposable elements as genomic diseases. Molecular bioSystems 5: 32-35.

[37] Harris CR, Normart R, Yang Q, Stevenson E, Haffty BG, et al. (2010) Association of nuclear localization of a long interspersed nuclear element-1 protein in breast tumors with poor prognostic outcomes. Genes & cancer 1: 115-124.

[38] Mirabello L, Savage SA, Korde L, Gadalla SM and Greene MH (2010) LINE-1 methylation is inherited in familial testicular cancer kindreds. BMC medical genetics 11: 77.

[39] Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, et al. (2010) L1 retrotransposition in neurons is modulated by MeCP2. Nature 468: 443-446.

[40] Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, et al. (2009) L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes & development 23: 1303-1312.

[41] Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, et al. (2009) L1 retrotransposition in human neural progenitor cells. Nature 460: 1127-1131.

[42] van den Hurk JA, Meij IC, Seleme MC, Kano H, Nikopoulos K, et al. (2007) L1 retrotransposition can occur early in human embryonic development. Human molecular genetics 16: 1587-1592.

[43] Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903-910.

[44] Cost GJ, Feng Q, Jacquier A and Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. The EMBO journal 21: 5899-5910.

[45] Dmitriev SE, Andreev DE, Terenin IM, Olovnikov IA, Prassolov VS, et al. (2007) Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5' untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated. Molecular and cellular biology 27: 4685-4697.

[46] Han JS (2010) Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mobile DNA 1: 15.

[47] Gasior SL, Wakeman TP, Xu B and Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. Journal of molecular biology 357: 1383-1393.

[48] Dombroski BA, Mathias SL, Nanthakumar E, Scott AF and Kazazian HH, Jr. (1991) Isolation of an active human transposable element. Science 254: 1805-1808.

[49] Johnson KJ, Springer NM, Bielinsky AK, Largaespada DA and Ross JA (2009) Developmental origins of cancer. Cancer research 69: 6375-6377.

[50] Holmes SE, Singer MF and Swergold GD (1992) Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. The Journal of biological chemistry 267: 19765-19768.

[51] Martin SL, Branciforte D, Keller D and Bain DL (2003) Trimeric structure for an essential protein in L1 retrotransposition. Proceedings of the National Academy of Sciences of the United States of America 100: 13815-13820.

[52] Januszyk K, Li PW, Villareal V, Branciforte D, Wu H, et al. (2007) Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. The Journal of biological chemistry 282: 24893-24904.

[53] Kolosha VO and Martin SL (2003) High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). The Journal of biological chemistry 278: 8112-8117.

[54] Khazina E and Weichenrieder O (2009) Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proceedings of the National Academy of Sciences of the United States of America 106: 731-736.

[55] Kulpa DA and Moran JV (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nature structural & molecular biology 13: 655-660.

[56] Hohjoh H and Singer MF (1997) Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. The EMBO journal 16: 6034-6043.

[57] Hohjoh H and Singer MF (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. The EMBO journal 15: 630-639.

[58] Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, et al. (2001) Human L1 retrotransposition: cis preference versus trans complementation. Molecular and cellular biology 21: 1429-1439.

[59] Kulpa DA and Moran JV (2005) Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Human molecular genetics 14: 3237-3248.

[60] Martin SL and Bushman FD (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Molecular and cellular biology 21: 467-475.

[61] Martin SL, Li J and Weisz JA (2000) Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. Journal of molecular biology 304: 11-20.

[62] Feng Q, Moran JV, Kazazian HH, Jr. and Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87: 905-916.

[63] Goodier JL, Mandal PK, Zhang L and Kazazian HH, Jr. (2010) Discrete subcellular partitioning of human retrotransposon RNAs despite a common mechanism of genome insertion. Human molecular genetics 19: 1712-1725.

[64] Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD and Gabriel A (1991) Reverse transcriptase encoded by a human transposable element. Science 254: 1808-1810.

[65] Wallace N, Wagstaff BJ, Deininger PL and Roy-Engel AM (2008) LINE-1 ORF1 protein enhances Alu SINE retrotransposition. Gene 419: 1-6.

[66] Huda A, Marino-Ramirez L and Jordan IK (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mobile DNA 1: 2.

[67] Huda A, Marino-Ramirez L, Landsman D and Jordan IK (2009) Repetitive DNA elements, nucleosome binding and human gene expression. Gene 436: 12-22.

[68] Nur I, Pascale E and Furano AV (1988) The left end of rat L1 (L1Rn, long interspersed repeated) DNA which is a CpG island can function as a promoter. Nucleic acids research 16: 9233-9251.

[69] Goyal R, Reinhardt R and Jeltsch A (2006) Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. Nucleic acids research 34: 1182-1188.

[70] Esteve PO, Chin HG, Benner J, Feehery GR, Samaranayake M, et al. (2009) Regulation of DNMT1 stability through SET7-mediated lysine methylation in mammalian cells. Proceedings of the National Academy of Sciences of the United States of America 106: 5076-5081.

[71] Glass JL, Fazzari MJ, Ferguson-Smith AC and Greally JM (2009) CG dinucleotide periodicities recognized by the Dnmt3a-Dnmt3L complex are distinctive at retroelements and imprinted domains. Mammalian genome : official journal of the International Mammalian Genome Society 20: 633-643.

[72] Brown KD and Robertson KD (2007) DNMT1 knockout delivers a strong blow to genome stability and cell viability. Nature genetics 39: 289-290.

[73] Kato Y, Kaneda M, Hata K, Kumaki K, Hisano M, et al. (2007) Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Human molecular genetics 16: 2272-2280.

[74] Gowher H, Liebert K, Hermann A, Xu G and Jeltsch A (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. The Journal of biological chemistry 280: 13341-13348.

[75] Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, et al. (2002) Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Molecular and cellular biology 22: 480-491.

[76] Jia D, Jurkowska RZ, Zhang X, Jeltsch A and Cheng X (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature 449: 248-251.

[77] Zilberman D, Gehring M, Tran RK, Ballinger T and Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nature genetics 39: 61-69.

[78] Gendrel AV and Colot V (2005) Arabidopsis epigenetics: when RNA meets chromatin. Current opinion in plant biology 8: 142-147.

[79] Ross JP, Suetake I, Tajima S and Molloy PL (2010) Recombinant mammalian DNA methyltransferase activity on model transcriptional gene silencing short RNA-DNA heteroduplex substrates. The Biochemical journal 432: 323-332.

[80] Ooi SK, Qiu C, Bernstein E, Li K, Jia D, et al. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature 448: 714-717.

[81] Hu JL, Zhou BO, Zhang RR, Zhang KL, Zhou JQ, et al. (2009) The N-terminus of histone H3 is required for de novo DNA methylation in chromatin. Proceedings of the National Academy of Sciences of the United States of America 106: 22187-22192.

[82] Garcia-Perez JL, Morell M, Scheys JO, Kulpa DA, Morell S, et al. (2010) Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature 466: 769-773.

[83] Chew YC, West JT, Kratzer SJ, Ilvarsonn AM, Eissenberg JC, et al. (2008) Biotinylation of histones represses transposable elements in human and mouse cells and cell lines and in Drosophila melanogaster. The Journal of nutrition 138: 2316-2322.

[84] Brunmeir R, Lagger S, Simboeck E, Sawicka A, Egger G, et al. (2010) Epigenetic regulation of a murine retrotransposon by a dual histone modification mark. PLoS genetics 6: e1000927.

[85] Zempleni J, Chew YC, Bao B, Pestinger V and Wijeratne SS (2009) Repression of transposable elements by histone biotinylation. The Journal of nutrition 139: 2389-2392.

[86] Questa JI, Walbot V and Casati P (2010) Mutator transposon activation after UV-B involves chromatin remodeling. Epigenetics : official journal of the DNA Methylation Society 5: 352-363.

[87] Sridhar VV, Kapoor A, Zhang K, Zhu J, Zhou T, et al. (2007) Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination. Nature 447: 735-738.

[88] Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, et al. (2001) Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Current biology : CB 11: 1017-1027.

[89] Sarot E, Payen-Groschene G, Bucheton A and Pelisson A (2004) Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. Genetics 166: 1313-1321.

[90] Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, et al. (2003) The small RNA profile during Drosophila melanogaster development. Developmental cell 5: 337-350.

[91] Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442: 203-207.

[92] Girard A, Sachidanandam R, Hannon GJ and Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442: 199-202.

[93] Grivna ST, Beyret E, Wang Z and Lin H (2006) A novel class of small RNAs in mouse spermatogenic cells. Genes & development 20: 1709-1714.

[94] Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, et al. (2006) Characterization of the piRNA complex from rat testes. Science 313: 363-367.

[95] Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, et al. (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313: 320-324.

[96] Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128: 1089-1103.

[97] Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, et al. (2009) Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell 137: 522-535.

[98] Li C, Vagin VV, Lee S, Xu J, Ma S, et al. (2009) Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell 137: 509-521.

[99] Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, et al. (2010) Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. Genes & development 24: 2493-2498.

[100] Qi H, Watanabe T, Ku HY, Liu N, Zhong M, et al. (2011) The Yb body, a major site for Piwi-associated RNA biogenesis and a gateway for Piwi expression and transport to the nucleus in somatic cells. The Journal of biological chemistry 286: 3789-3797.

[101] Haase AD, Fenoglio S, Muerdter F, Guzzardo PM, Czech B, et al. (2010) Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila. Genes & development 24: 2499-2504.

[102] Olivieri D, Sykora MM, Sachidanandam R, Mechtler K and Brennecke J (2010) An in vivo RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in Drosophila. The EMBO journal 29: 3301-3317.

[103] Szakmary A, Reedy M, Qi H and Lin H (2009) The Yb protein defines a novel organelle and regulates male germline stem cell self-renewal in Drosophila melanogaster. The Journal of cell biology 185: 613-627.

[104] Pane A, Wehr K and Schupbach T (2007) zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. Developmental cell 12: 851-862.

[105] Thomson T, Liu N, Arkov A, Lehmann R and Lasko P (2008) Isolation of new polar granule components in Drosophila reveals P body and ER associated proteins. Mechanisms of development 125: 865-873.

[106] Lau NC, Robine N, Martin R, Chung WJ, Niki Y, et al. (2009) Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome research 19: 1776-1785.

[107] Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, et al. (2008) An epigenetic role for maternally inherited piRNAs in transposon silencing. Science 322: 1387-1392.

[108] Lim AK, Tao L and Kai T (2009) piRNAs mediate posttranscriptional retroelement silencing and localization to pi-bodies in the Drosophila germline. The Journal of cell biology 186: 333-342.

[109] Rouget C, Papin C, Boureux A, Meunier AC, Franco B, et al. (2010) Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. Nature 467: 1128-1132.

[110] Trelogan SA and Martin SL (1995) Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis.

Proceedings of the National Academy of Sciences of the United States of America 92: 1520-1524.

[111] Aravin AA, van der Heijden GW, Castaneda J, Vagin VV, Hannon GJ, et al. (2009) Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. PLoS genetics 5: e1000764.

[112] Soper SF, van der Heijden GW, Hardiman TC, Goodheart M, Martin SL, et al. (2008) Mouse maelstrom, a component of nuage, is essential for spermatogenesis and transposon repression in meiosis. Developmental cell 15: 285-297.

[113] Ma L, Buchold GM, Greenbaum MP, Roy A, Burns KH, et al. (2009) GASZ is essential for male meiosis and suppression of retrotransposon expression in the male germline. PLoS genetics 5: e1000635.

[114] Frost RJ, Hamra FK, Richardson JA, Qi X, Bassel-Duby R, et al. (2010) MOV10L1 is necessary for protection of spermatocytes against retrotransposons by Piwi-interacting RNAs. Proceedings of the National Academy of Sciences of the United States of America 107: 11847-11852.

[115] Shoji M, Tanaka T, Hosokawa M, Reuter M, Stark A, et al. (2009) The TDRD9-MIWI2 complex is essential for piRNA-mediated retrotransposon silencing in the mouse male germline. Developmental cell 17: 775-787.

[116] Kuramochi-Miyagawa S, Kimura T, Yomogida K, Kuroiwa A, Tadokoro Y, et al. (2001) Two mouse piwi-related genes: miwi and mili. Mechanisms of development 108: 121-133.

[117] Zheng K, Xiol J, Reuter M, Eckardt S, Leu NA, et al. (2010) Mouse MOV10L1 associates with Piwi proteins and is an essential component of the Piwi-interacting RNA (piRNA) pathway. Proceedings of the National Academy of Sciences of the United States of America 107: 11841-11846.

[118] Yan W, Rajkovic A, Viveiros MM, Burns KH, Eppig JJ, et al. (2002) Identification of Gasz, an evolutionarily conserved gene expressed exclusively in germ cells and encoding a protein with four ankyrin repeats, a sterile-alpha motif, and a basic leucine zipper. Molecular endocrinology 16: 1168-1184.

[119] Reuter M, Chuma S, Tanaka T, Franz T, Stark A, et al. (2009) Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile. Nature structural & molecular biology 16: 639-646.

[120] Halic M and Moazed D (2010) Dicer-independent primal RNAs trigger RNAi and heterochromatin formation. Cell 140: 504-516.

[121] Wang J, Saxe JP, Tanaka T, Chuma S and Lin H (2009) Mili interacts with tudor domain-containing protein 1 in regulating spermatogenesis. Current biology : CB 19: 640-644.

[122] Kuramochi-Miyagawa S, Kimura T, Ijiri TW, Isobe T, Asada N, et al. (2004) Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Development 131: 839-849.

[123] Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, et al. (2007) MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Developmental cell 12: 503-514.

[124] Teixeira D and Parker R (2007) Analysis of P-body assembly in Saccharomyces cerevisiae. Molecular biology of the cell 18: 2274-2287.

[125] Ma JB, Yuan YR, Meister G, Pei Y, Tuschl T, et al. (2005) Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature 434: 666-670.

[126] Parker JS, Roe SM and Barford D (2005) Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. Nature 434: 663-666.

[127] Parker R and Sheth U (2007) P bodies and the control of mRNA translation and degradation. Molecular cell 25: 635-646.

[128] Wassenegger M (2000) RNA-directed DNA methylation. Plant molecular biology 43: 203-220.

[129] Gao Z, Liu HL, Daxinger L, Pontes O, He X, et al. (2010) An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. Nature 465: 106-109.

[130] Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Takamatsu K, Chuma S, et al. (2010) MVH in piRNA processing and gene silencing of retrotransposons. Genes & development 24: 887-892.

[131] Matzke M, Kanno T, Huettel B, Daxinger L and Matzke AJ (2007) Targets of RNA-directed DNA methylation. Current opinion in plant biology 10: 512-519.

[132] Mathieu O and Bender J (2004) RNA-directed DNA methylation. Journal of cell science 117: 4881-4888.

[133] Vagin VV, Wohlschlegel J, Qu J, Jonsson Z, Huang X, et al. (2009) Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. Genes & development 23: 1749-1762.

[134] Tanaka SS, Toyooka Y, Akasu R, Katoh-Fukui Y, Nakahara Y, et al. (2000) The mouse homolog of Drosophila Vasa is required for the development of male germ cells. Genes & development 14: 841-853.

[135] Chen C, Jin J, James DA, Adams-Cioaba MA, Park JG, et al. (2009) Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. Proceedings of the National Academy of Sciences of the United States of America 106: 20336-20341.

[136] Biessmann H, Valgeirsdottir K, Lofsky A, Chin C, Ginther B, et al. (1992) HeT-A, a transposable element specifically involved in "healing" broken chromosome ends in Drosophila melanogaster. Molecular and cellular biology 12: 3910-3918.

[137] Fugmann SD (2010) The origins of the Rag genes--from transposition to V(D)J recombination. Seminars in immunology 22: 10-16.

[138] Levis RW, Ganesan R, Houtchens K, Tolar LA and Sheen FM (1993) Transposons in place of telomeric repeats at a Drosophila telomere. Cell 75: 1083-1093.

[139] Chalker DL (2009) Transposons that clean up after themselves. Genome biology 10: 224.

[140] McVicker G and Green P (2010) Genomic signatures of germline gene expression. Genome research 20: 1503-1511.

[141] Finnegan DJ (2009) Genome dynamics: transposition and the single cell. Current biology : CB 19: R555-558.

[142] Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, et al. (2009) A functional role for transposases in a large eukaryotic genome. Science 324: 935-938.

[143] Kvikstad EM and Makova KD (2010) The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. Genome research 20: 600-613.

[144] Beraldi R, Pittoggi C, Sciamanna I, Mattei E and Spadafora C (2006) Expression of LINE-1 retroposons is essential for murine preimplantation development. Molecular reproduction and development 73: 279-287.

[145] Dupuy AJ (2010) Current applications of transposons in mouse genetics. Methods in enzymology 477: 53-70.

[146] Sciamanna I, Barberi L, Martire A, Pittoggi C, Beraldi R, et al. (2003) Sperm endogenous reverse transcriptase as mediator of new genetic information. Biochemical and biophysical research communications 312: 1039-1046.

[147] Pittoggi C, Sciamanna I, Mattei E, Beraldi R, Lobascio AM, et al. (2003) Role of endogenous reverse transcriptase in murine early embryo development. Molecular reproduction and development 66: 225-236.

[148] Rubin GM and Spradling AC (1982) Genetic transformation of Drosophila with transposable element vectors. Science 218: 348-353.

[149] Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, et al. (1999) The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics 153: 135-177.

[150] Spradling AC and Rubin GM (1982) Transposition of cloned P elements into Drosophila germ line chromosomes. Science 218: 341-347.

[151] Lin H and Spradling AC (1997) A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. Development 124: 2463-2476.

[152] Ivics Z, Hackett PB, Plasterk RH and Izsvak Z (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91: 501-510.

[153] Dupuy AJ, Akagi K, Largaespada DA, Copeland NG and Jenkins NA (2005) Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. Nature 436: 221-226.

[154] Castaneda J, Genzor P, van der Heijden GW, Sarkeshik A, Yates JR, 3rd, et al. (2014) Reduced pachytene piRNAs and translation underlie spermiogenic arrest in Maelstrom mutant mice. The EMBO journal 33: 1999-2019.

[155] Dupuy AJ, Rogers LM, Kim J, Nannapaneni K, Starr TK, et al. (2009) A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. Cancer research 69: 8150-8156.

[156] Largaespada DA (2009) Transposon-mediated mutagenesis of somatic cells in the mouse for cancer gene identification. Methods 49: 282-286.

[157] Kitada K, Keng VW, Takeda J and Horie K (2009) Generating mutant rats using the Sleeping Beauty transposon system. Methods 49: 236-242.

[158] Liu L, Liu H, Visner G and Fletcher BS (2006) Sleeping Beauty-mediated eNOS gene therapy attenuates monocrotaline-induced pulmonary hypertension in rats. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 20: 2594-2596.

[159] Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, et al. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 453: 539-543.

[160] Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, et al. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 453: 534-538.

[161] Clegg NJ, Frost DM, Larkin MK, Subrahmanyan L, Bryant Z, et al. (1997) maelstrom is required for an early step in the establishment of Drosophila oocyte polarity: posterior localization of grk mRNA. Development 124: 4661-4671.

[162] Eddy EM (1975) Germ plasm and the differentiation of the germ cell line. International review of cytology 43: 229-280.

[163] Voronina E, Seydoux G, Sassone-Corsi P and Nagamori I (2011) RNA granules in germ cells. Cold Spring Harbor perspectives in biology 3.

[164] Clegg NJ, Findley SD, Mahowald AP and Ruohola-Baker H (2001) Maelstrom is required to position the MTOC in stage 2-6 Drosophila oocytes. Development genes and evolution 211: 44-48.

[165] Findley SD, Tamanaha M, Clegg NJ and Ruohola-Baker H (2003) Maelstrom, a Drosophila spindle-class gene, encodes a protein that colocalizes with Vasa and RDE1/AGO1 homolog, Aubergine, in nuage. Development 130: 859-871.

[166] Pek JW, Lim AK and Kai T (2009) Drosophila maelstrom ensures proper germline stem cell lineage differentiation by repressing microRNA-7. Developmental cell 17: 417-424.

[167] Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, et al. (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes & development 20: 1732-1743.

[168] Aravin AA, Hannon GJ and Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318: 761-764.

[169] Zhang D, Xiong H, Shan J, Xia X and Trudeau VL (2008) Functional insight into Maelstrom in the germline piRNA pathway: a unique domain homologous to the DnaQ-H 3'-5' exonuclease, its lineage-specific expansion/loss and evolutionarily active site switch. Biology direct 3: 48.

[170] Majorek KA, Dunin-Horkawicz S, Steczkiewicz K, Muszewska A, Nowotny M, et al. (2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic acids research 42: 4160-4179.

[171] Sato K, Nishida KM, Shibuya A, Siomi MC and Siomi H (2011) Maelstrom coordinates microtubule organization during Drosophila oogenesis through interaction with components of the MTOC. Genes & development 25: 2361-2373.

[172] Sienski G, Donertas D and Brennecke J (2012) Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell 151: 964-980.

[173] Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proceedings of the National Academy of Sciences of the United States of America 103: 17600-17601.

[174] Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, et al. (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nature genetics 31: 159-165.

[175] Mizuuchi M, Baker TA and Mizuuchi K (1991) DNase protection analysis of the stable synaptic complexes involved in Mu transposition. Proceedings of the National Academy of Sciences of the United States of America 88: 9031-9035.

[176] Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution. Science 303: 1626-1632.

[177] Muotri AR, Marchetto MC, Coufal NG and Gage FH (2007) The necessary junk: new functions for transposable elements. Human molecular genetics 16 Spec No. 2: R159-167.

[178] Vourekas A, Zheng Q, Alexiou P, Maragkakis M, Kirino Y, et al. (2012) Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nature structural & molecular biology 19: 773-781.

[179] Li XZ, Roy CK, Dong X, Bolcun-Filas E, Wang J, et al. (2013) An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Molecular cell 50: 67-81.

[180] van der Heijden GW, Castaneda J and Bortvin A (2010) Bodies of evidence - compartmentalization of the piRNA pathway in mouse fetal prospermatogonia. Current opinion in cell biology 22: 752-757.

[181] Malki S, van der Heijden GW, O'Donnell KA, Martin SL and Bortvin A (2014) A role for retrotransposon LINE-1 in fetal oocyte attrition in mice. Developmental cell 29: 521-533.

[182] Tomizawa T, Kigawa, T., Sato, M., Koshiba, S., Inoue, M., Kamatari, Y.O., Yokoyama, S., RIKEN Structural Genomics/Proteomics Initiative (2005) Solution

structure of the HMG box like domain from human hypothetical protein FLJ14904. RCSC PDB: Protein Data Bank. pp. 2cto: structure of human MAEL HMG-box.

[183] Johns EW (1982) The HMG chromosomal proteins. London ; New York: Academic Press. viii, 251 p. p.

[184] Bustin M (2001) Revised nomenclature for high mobility group (HMG) chromosomal proteins. Trends in biochemical sciences 26: 152-153.

[185] Huth JR, Bewley CA, Nissen MS, Evans JN, Reeves R, et al. (1997) The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. Nature structural biology 4: 657-665.

[186] Aravind L and Landsman D (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic acids research 26: 4413-4421.

[187] Reeves R and Nissen MS (1990) The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. The Journal of biological chemistry 265: 8573-8582.

[188] Fonfria-Subiros E, Acosta-Reyes F, Saperas N, Pous J, Subirana JA, et al. (2012) Crystal structure of a complex of DNA with one AT-hook of HMGA1. PloS one 7: e37120.

[189] Fedele M and Fusco A (2010) HMGA and cancer. Biochimica et biophysica acta 1799: 48-54.

[190] Sgarra R, Zammitti S, Lo Sardo A, Maurizio E, Arnoldo L, et al. (2010) HMGA molecular network: From transcriptional regulation to chromatin remodeling. Biochimica et biophysica acta 1799: 37-47.

[191] Ueda T, Catez F, Gerlitz G and Bustin M (2008) Delineation of the protein module that anchors HMGN proteins to nucleosomes in the chromatin of living cells. Molecular and cellular biology 28: 2872-2883.

[192] Ueda T, Postnikov YV and Bustin M (2006) Distinct domains in high mobility group N variants modulate specific chromatin modifications. The Journal of biological chemistry 281: 10182-10187.

[193] Rochman M, Postnikov Y, Correll S, Malicet C, Wincovitch S, et al. (2009) The interaction of NSBP1/HMGN5 with nucleosomes in euchromatin counteracts linker histone-mediated chromatin compaction and modulates transcription. Molecular cell 35: 642-656.

[194] Catez F, Yang H, Tracey KJ, Reeves R, Misteli T, et al. (2004) Network of dynamic interactions between histone H1 and high-mobility-group proteins in chromatin. Molecular and cellular biology 24: 4321-4328.

[195] Furusawa T, Lim JH, Catez F, Birger Y, Mackem S, et al. (2006) Down-regulation of nucleosomal binding protein HMGN1 expression during embryogenesis modulates Sox9 expression in chondrocytes. Molecular and cellular biology 26: 592-604.

[196] Postnikov Y and Bustin M (2010) Regulation of chromatin structure and function by HMGN proteins. Biochimica et biophysica acta 1799: 62-68.

[197] Furusawa T and Cherukuri S (2010) Developmental function of HMGN proteins. Biochimica et biophysica acta 1799: 69-73.

[198] Gerlitz G (2010) HMGNs, DNA repair and cancer. Biochimica et biophysica acta 1799: 80-85.

[199] Pogna EA, Clayton AL and Mahadevan LC (2010) Signalling to chromatin through post-translational modifications of HMGN. Biochimica et biophysica acta 1799: 93-100.

[200] Teo SH, Grasser KD, Hardman CH, Broadhurst RW, Laue ED, et al. (1995) Two mutations in the HMG-box with very different structural consequences provide insights into the nature of binding to four-way junction DNA. The EMBO journal 14: 3844-3853.

[201] Racca JD, Chen YS, Maloy JD, Wickramasinghe N, Phillips NB, et al. (2014) Structure-function relationships in human testis-determining factor SRY: an aromatic buttress underlies the specific DNA-bending surface of a high mobility group (HMG) box. The Journal of biological chemistry 289: 32410-32429.

[202] Stros M (2001) Two mutations of basic residues within the N-terminus of HMG-1 B domain with different effects on DNA supercoiling and binding to bent DNA. Biochemistry 40: 4769-4779.

[203] Giese K, Amsterdam A and Grosschedl R (1991) DNA-binding properties of the HMG domain of the lymphoid-specific transcriptional regulator LEF-1. Genes & development 5: 2567-2578.

[204] Klass J, Murphy FVt, Fouts S, Serenil M, Changela A, et al. (2003) The role of intercalating residues in chromosomal high-mobility-group protein DNA binding, bending and specificity. Nucleic acids research 31: 2852-2864.

[205] Thomsen MS, Franssen L, Launholt D, Fojan P and Grasser KD (2004) Interactions of the basic N-terminal and the acidic C-terminal domains of the maize chromosomal HMGB1 protein. Biochemistry 43: 8029-8037.

[206] Grasser M, Christensen JM, Peterhansel C and Grasser KD (2007) Basic and acidic regions flanking the HMG-box domain of maize HMGB1 and HMGB5 modulate the stimulatory effect on the DNA binding of transcription factor Dof2. Biochemistry 46: 6375-6382.

[207] Lnenicek-Allen M, Read CM and Crane-Robinson C (1996) The DNA bend angle and binding affinity of an HMG box increased by the presence of short terminal arms. Nucleic acids research 24: 1047-1051.

[208] Dragan AI, Read CM, Makeyeva EN, Milgotina EI, Churchill ME, et al. (2004) DNA binding and bending by HMG boxes: energetic determinants of specificity. Journal of molecular biology 343: 371-393.

[209] Desclozeaux M, Poulat F, de Santa Barbara P, Capony JP, Turowski P, et al. (1998) Phosphorylation of an N-terminal motif enhances DNA-binding activity of the human SRY protein. The Journal of biological chemistry 273: 7988-7995.

[210] Sterner R, Vidali G and Allfrey VG (1979) Studies of acetylation and deacetylation in high mobility group proteins. Identification of the sites of acetylation in HMG-1. The Journal of biological chemistry 254: 11577-11583.

[211] Bonaldi T, Talamo F, Scaffidi P, Ferrera D, Porto A, et al. (2003) Monocytic cells hyperacetylate chromatin protein HMGB1 to redirect it towards secretion. The EMBO journal 22: 5551-5560.

[212] Youn JH and Shin JS (2006) Nucleocytoplasmic shuttling of HMGB1 is regulated by phosphorylation that redirects it toward secretion. Journal of immunology 177: 7889-7897.

[213] Ito I, Fukazawa J and Yoshida M (2007) Post-translational methylation of high mobility group box 1 (HMGB1) causes its cytoplasmic localization in neutrophils. The Journal of biological chemistry 282: 16336-16344.

[214] Stros M (2010) HMGB proteins: interactions with DNA and chromatin. Biochimica et biophysica acta 1799: 101-113.

[215] Park S and Lippard SJ (2012) Binding interaction of HMGB4 with cisplatin-modified DNA. Biochemistry 51: 6728-6737.

[216] Pil PM and Lippard SJ (1992) Specific binding of chromosomal protein HMG1 to DNA damaged by the anticancer drug cisplatin. Science 256: 234-237.

[217] Bianchi ME, Beltrame M and Paonessa G (1989) Specific recognition of cruciform DNA by nuclear protein HMG1. Science 243: 1056-1059.

[218] Travers A (2000) Recognition of distorted DNA structures by HMG domains. Current opinion in structural biology 10: 102-109.

[219] Murphy EC, Zhurkin VB, Louis JM, Cornilescu G and Clore GM (2001) Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. Journal of molecular biology 312: 481-499.

[220] Mertin S, McDowall SG and Harley VR (1999) The DNA-binding specificity of SOX9 and other SOX proteins. Nucleic acids research 27: 1359-1364.

[221] Ohndorf UM, Rould MA, He Q, Pabo CO and Lippard SJ (1999) Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. Nature 399: 708-712.

[222] Jung Y and Lippard SJ (2003) Nature of full-length HMGB1 binding to cisplatin-modified DNA. Biochemistry 42: 2664-2671.

[223] Webb M and Thomas JO (1999) Structure-specific binding of the two tandem HMG boxes of HMG1 to four-way junction DNA is mediated by the A domain. Journal of molecular biology 294: 373-387.

[224] Ferrari S, Harley VR, Pontiggia A, Goodfellow PN, Lovell-Badge R, et al. (1992) SRY, like HMG1, recognizes sharp angles in DNA. The EMBO journal 11: 4497-4506.

[225] JR Po, Norman DG, Bramham J, Bianchi ME and Lilley DM (1998) HMG box proteins bind to four-way DNA junctions in their open conformation. The EMBO journal 17: 817-826.

[226] Zlatanova J and van Holde K (1998) Binding to four-way junction DNA: a common property of architectural proteins? FASEB journal : official publication of the Federation of American Societies for Experimental Biology 12: 421-431.

[227] Clegg RM, Murchie AI and Lilley DM (1994) The solution structure of the four-way DNA junction at low-salt conditions: a fluorescence resonance energy transfer analysis. Biophysical journal 66: 99-109.

[228] Eichman BF, Ortiz-Lombardia M, Aymami J, Coll M and Ho PS (2002) The inherent properties of DNA four-way junctions: comparing the crystal structures of holliday junctions. Journal of molecular biology 320: 1037-1051.

[229] Liberi G and Foiani M (2010) The double life of Holliday junctions. Cell research 20: 611-613.

[230] Holliday R (1964) The Induction of Mitotic Recombination by Mitomycin C in Ustilago and Saccharomyces. Genetics 50: 323-335.

[231] Bzymek M, Thayer NH, Oh SD, Kleckner N and Hunter N (2010) Double Holliday junctions are intermediates of DNA break repair. Nature 464: 937-941.

[232] Brazda V, Laister RC, Jagelska EB and Arrowsmith C (2011) Cruciform structures are a common DNA feature important for regulating biological processes. BMC molecular biology 12: 33.

[233] Hsu T, King DL, LaBonne C and Kafatos FC (1993) A Drosophila single-strand DNA/RNA-binding factor contains a high-mobility-group box and is enriched in the nucleolus. Proceedings of the National Academy of Sciences of the United States of America 90: 6488-6492.

[234] Copenhaver GP, Putnam CD, Denton ML and Pikaard CS (1994) The RNA polymerase I transcription factor UBF is a sequence-tolerant HMG-box protein that can recognize structured nucleic acids. Nucleic acids research 22: 2651-2657.

[235] Arimondo PB, Gelus N, Hamy F, Payet D, Travers A, et al. (2000) The chromosomal protein HMG-D binds to the TAR and RBE RNA of HIV-1. FEBS letters 485: 47-52.

[236] Veretnik S and Gribskov M (1999) RNA binding domain of HDV antigen is homologous to the HMG box of SRY. Archives of virology 144: 1139-1158.

[237] Bell AJ, Jr., Chauhan S, Woodson SA and Kallenbach NR (2008) Interactions of recombinant HMGB proteins with branched RNA substrates. Biochemical and biophysical research communications 377: 262-267.

[238] Butcher SE and Pyle AM (2011) The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. Accounts of chemical research 44: 1302-1311.

[239] Laing C and Schlick T (2009) Analysis of four-way junctions in RNA structures. Journal of molecular biology 390: 547-559.

[240] Hohng S, Wilson TJ, Tan E, Clegg RM, Lilley DM, et al. (2004) Conformational flexibility of four-way junctions in RNA. Journal of molecular biology 336: 69-79.

[241] Yanai H, Chiba S, Ban T, Nakaima Y, Onoe T, et al. (2011) Suppression of immune responses by nonimmunogenic oligodeoxynucleotides with high affinity for high-mobility group box proteins (HMGBs). Proceedings of the National Academy of Sciences of the United States of America 108: 11542-11547.

[242] Yanai H, Ban T, Wang Z, Choi MK, Kawamura T, et al. (2009) HMGB proteins function as universal sentinels for nucleic-acid-mediated innate immune responses. Nature 462: 99-103.

[243] Woolhouse ME and Adair K (2013) Ecological and taxonomic variation among human RNA viruses. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology 58: 344-345.

[244] Laudet V, Stehelin D and Clevers H (1993) Ancestry and diversity of the HMG box superfamily. Nucleic acids research 21: 2493-2501.

[245] Stros M, Launholt D and Grasser KD (2007) The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cellular and molecular life sciences : CMLS 64: 2590-2606.

[246] Tamura K, Stecher G, Peterson D, Filipski A and Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular biology and evolution 30: 2725-2729.

[247] Jones DT, Taylor WR and Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences : CABIOS 8: 275-282.

[248] Kim DE, Chivian D and Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic acids research 32: W526-531.

[249] Waterhouse AM, Procter JB, Martin DM, Clamp M and Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

[250] Moore S Round-the-horn site-directed mutagenesis. OpenWetWare.

[251] Greenfield NJ (2006) Using circular dichroism spectra to estimate protein secondary structure. Nature protocols 1: 2876-2890.

[252] Chen Z and Ruffner DE (1996) Modified crush-and-soak method for recovering oligodeoxynucleotides from polyacrylamide gel. BioTechniques 21: 820-822.

[253] Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research 31: 3406-3415.

[254] Tinoco I, Jr. and Bustamante C (1999) How RNA folds. Journal of molecular biology 293: 271-281.

[255] Ryder SP, Recht MI and Williamson JR (2008) Quantitative analysis of protein-RNA interactions by gel mobility shift. Methods in molecular biology 488: 99-115.

[256] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome biology 9: R137.

[257] Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

[258] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell 38: 576-589.

[259] Thorvaldsdottir H, Robinson JT and Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics 14: 178-192.

[260] Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nature biotechnology 29: 24-26.

[261] Zappulla DC and Cech TR (2004) Yeast telomerase RNA: a flexible scaffold for protein subunits. Proceedings of the National Academy of Sciences of the United States of America 101: 10024-10029.

[262] Bernhart SH, Hofacker IL, Will S, Gruber AR and Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. BMC bioinformatics 9: 474.

[263] Xue B, Dunbrack RL, Williams RW, Dunker AK and Uversky VN (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochimica et biophysica acta 1804: 996-1010.

[264] Blom N, Gammeltoft S and Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of molecular biology 294: 1351-1362.

[265] Luscombe NM, Laskowski RA and Thornton JM (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic acids research 29: 2860-2874.

[266] Richardson JS and Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices. Science 240: 1648-1652.

[267] Pace CN and Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. Biophysical journal 75: 422-427.

[268] Khoury GA, Baliban RC and Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Scientific reports 1.

[269] George RA and Heringa J (2002) An analysis of protein domain linkers: their classification and role in protein folding. Protein engineering 15: 871-879.

[270] Gerstein M, Lesk AM and Chothia C (1994) Structural mechanisms for domain movements in proteins. Biochemistry 33: 6739-6749.

[271] Vuzman D and Levy Y (2012) Intrinsically disordered regions as affinity tuners in protein-DNA interactions. Molecular bioSystems 8: 47-57.

[272] Nishi H, Fong JH, Chang C, Teichmann SA and Panchenko AR (2013) Regulation of protein-protein binding by coupling between phosphorylation and intrinsic disorder: analysis of human protein complexes. Molecular bioSystems 9: 1620-1626.

[273] Pek JW, Ng BF and Kai T (2012) Polo-mediated phosphorylation of Maelstrom regulates oocyte determination during oogenesis in Drosophila. Development 139: 4505-4513.

[274] Kramer G, Kudlicki W, McCarthy D, Tsalkova T, Simmons D, et al. (1999) N-terminal and C-terminal modifications affect folding, release from the ribosomes and stability of in vitro synthesized proteins. The international journal of biochemistry & cell biology 31: 231-241.

[275] Walsh CT, Garneau-Tsodikova S and Gatto GJ, Jr. (2005) Protein posttranslational modifications: the chemistry of proteome diversifications. Angewandte Chemie 44: 7342-7372.

[276] Uversky VN (2013) The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. FEBS letters 587: 1891-1901.

[277] Kramer G, Ramachandiran V and Hardesty B (2001) Cotranslational folding--omnia mea mecum porto? The international journal of biochemistry & cell biology 33: 541-553.

[278] Ainavarapu SR, Brujic J, Huang HH, Wiita AP, Lu H, et al. (2007) Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. Biophysical journal 92: 225-233.

[279] Carrion-Vazquez M, Marszalek PE, Oberhauser AF and Fernandez JM (1999) Atomic force microscopy captures length phenotypes in single proteins. Proceedings of the National Academy of Sciences of the United States of America 96: 11288-11292.

[280] Stagno JR, Ma B, Li J, Altieri AS, Byrd RA, et al. (2012) Crystal structure of a plectonemic RNA supercoil. Nature communications 3: 901.

[281] Moult J, Fidelis K, Kryshtafovych A, Schwede T and Tramontano A (2014) Critical assessment of methods of protein structure prediction (CASP)--round x. Proteins 82 Suppl 2: 1-6.

[282] Cheng Y, Patel, D.J. (2005) Structural basis for 3'-end specific recognition of histone mRNA stem-loop by 3'-exonuclease, a human nuclease that also targets siRNA. RCSC PDB: Protein Data Bank.

[283] Singleton MR, Dillingham MS, Gaudier M, Kowalczykowski SC and Wigley DB (2004) Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. Nature 432: 187-193.

[284] Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, et al. (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science 343: 1247997.

[285] Zuo Y, Vincent HA, Zhang J, Wang Y, Deutscher MP, et al. (2006) Structural basis for processivity and single-strand specificity of RNase II. Molecular cell 24: 149-156.

[286] Phillips GN, Jr. (2009) Describing protein conformational ensembles: beyond static snapshots. F1000 biology reports 1: 38.

[287] Kumar S and Hedges SB (2011) TimeTree2: species divergence times on the iPhone. Bioinformatics 27: 2023-2024.

[288] Hedges SB, Dudley J and Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22: 2971-2972.

[289] King CY and Weiss MA (1993) The SRY high-mobility-group box recognizes DNA by partial intercalation in the minor groove: a topological mechanism of sequence specificity. Proceedings of the National Academy of Sciences of the United States of America 90: 11990-11994.

[290] McCauley MJ, Zimmerman J, Maher LJ, 3rd and Williams MC (2007) HMGB binding to DNA: single and double box motifs. Journal of molecular biology 374: 993-1004.

[291] Jaouen S, de Koning L, Gaillard C, Muselikova-Polanska E, Stros M, et al. (2005) Determinants of specific binding of HMGB1 protein to hemicatenated DNA loops. Journal of molecular biology 353: 822-837.

[292] Yoshioka K, Saito K, Tanabe T, Yamamoto A, Ando Y, et al. (1999) Differences in DNA recognition and conformational change activity between boxes A and B in HMG2 protein. Biochemistry 38: 589-595.

[293] Thomas JO and Travers AA (2001) HMG1 and 2, and related 'architectural' DNA-binding proteins. Trends in biochemical sciences 26: 167-174.

[294] West SM, Rohs R, Mann RS and Honig B (2010) Electrostatic interactions between arginines and the minor groove in the nucleosome. Journal of biomolecular structure & dynamics 27: 861-866.

[295] Rohs R, Jin X, West SM, Joshi R, Honig B, et al. (2010) Origins of specificity in protein-DNA recognition. Annual review of biochemistry 79: 233-269.

[296] Calnan BJ, Tidor B, Biancalana S, Hudson D and Frankel AD (1991) Arginine-mediated RNA recognition: the arginine fork. Science 252: 1167-1171.

[297] Bewley CA, Gronenborn AM and Clore GM (1998) Minor groove-binding architectural proteins: structure, function, and DNA recognition. Annual review of biophysics and biomolecular structure 27: 105-131.

[298] Hellman LM and Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nature protocols 2: 1849-1861.

[299] Yakhnin AV, Yakhnin H and Babitzke P (2012) Gel mobility shift assays to detect protein-RNA interactions. Methods in molecular biology 905: 201-211.

[300] Harley VR, Lovell-Badge R and Goodfellow PN (1994) Definition of a consensus DNA binding site for SRY. Nucleic acids research 22: 1500-1501.

[301] Pontiggia A, Rimini R, Harley VR, Goodfellow PN, Lovell-Badge R, et al. (1994) Sex-reversing mutations affect the architecture of SRY-DNA complexes. The EMBO journal 13: 6115-6124.

[302] van Buuren BN, Hermann T, Wijmenga SS and Westhof E (2002) Brownian-dynamics simulations of metal-ion binding to four-way junctions. Nucleic acids research 30: 507-514.

[303] Yu J, Ha T and Schulten K (2004) Conformational model of the Holliday junction transition deduced from molecular dynamics simulations. Nucleic acids research 32: 6683-6695.

[304] Nowakowski J, Shim PJ, Stout CD and Joyce GF (2000) Alternative conformations of a nucleic acid four-way junction. Journal of molecular biology 300: 93-102.

[305] Gaillard C and Strauss F (2000) High affinity binding of proteins HMG1 and HMG2 to semicatenated DNA loops. BMC molecular biology 1: 1.

[306] Ohno T, Umeda S, Hamasaki N and Kang D (2000) Binding of human mitochondrial transcription factor A, an HMG box protein, to a four-way DNA junction. Biochemical and biophysical research communications 271: 492-498.

[307] He Q, Ohndorf UM and Lippard SJ (2000) Intercalating residues determine the mode of HMG1 domains A and B binding to cisplatin-modified DNA. Biochemistry 39: 14426-14435.

[308] Uhlenbeck OC, Pardi A and Feigon J (1997) RNA structure comes of age. Cell 90: 833-840.

[309] Svoboda P and Di Cara A (2006) Hairpin RNA: a secondary structure of primary importance. Cellular and molecular life sciences : CMLS 63: 901-908.

[310] Watson JD (2008) Molecular biology of the gene. San Francisco
Cold Spring Harbor, N.Y.: Pearson/Benjamin Cummings ;
Cold Spring Harbor Laboratory Press. xxxii, 841 p. p.

[311] Serganov A and Patel DJ (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. Nature reviews Genetics 8: 776-790.

[312] Lazinski D, Grzadzielska E and Das A (1989) Sequence-specific recognition of RNA hairpins by bacteriophage antiterminators requires a conserved arginine-rich motif. Cell 59: 207-218.

[313] Turnage MA, Brewer-Jensen P, Bai WL and Searles LL (2000) Arginine-rich regions mediate the RNA binding and regulatory activities of the protein encoded by the Drosophila melanogaster suppressor of sable gene. Molecular and cellular biology 20: 8198-8208.

[314] Borders CL, Jr., Broadwater JA, Bekeny PA, Salmon JE, Lee AS, et al. (1994) A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens. Protein science : a publication of the Protein Society 3: 541-548.

[315] Shimoni L and Glusker JP (1995) Hydrogen bonding motifs of protein side chains: descriptions of binding of arginine and amide groups. Protein science : a publication of the Protein Society 4: 65-74.

[316] Word JM, Lovell SC, Richardson JS and Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of molecular biology 285: 1735-1747.

[317] Gary JD and Clarke S (1998) RNA and protein interactions modulated by protein arginine methylation. Progress in nucleic acid research and molecular biology 61: 65-131.

[318] Bayer TS, Booth LN, Knudsen SM and Ellington AD (2005) Arginine-rich motifs present multiple interfaces for specific binding by RNA. Rna 11: 1848-1857.

[319] Laing C, Wen D, Wang JT and Schlick T (2012) Predicting coaxial helical stacking in RNA junctions. Nucleic acids research 40: 487-498.

[320] Severynse DM, Hutchison CA, 3rd and Edgell MH (1992) Identification of transcriptional regulatory activity within the 5' A-type monomer sequence of the mouse LINE-1 retroposon. Mammalian genome : official journal of the International Mammalian Genome Society 2: 41-50.

[321] Schichman SA, Adey NB, Edgell MH and Hutchison CA, 3rd (1993) L1 A-monomer tandem arrays have expanded during the course of mouse L1 evolution. Molecular biology and evolution 10: 552-570.

[322] Parsch J, Braverman JM and Stephan W (2000) Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics 154: 909-921.

[323] Sookdeo A, Hepp CM, McClure MA and Boissinot S (2013) Revisiting the evolution of mouse LINE-1 in the genomic era. Mobile DNA 4: 3.

[324] Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, et al. (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Molecular cell 54: 887-900.

[325] Soullier S, Jay P, Poulat F, Vanacker JM, Berta P, et al. (1999) Diversification pattern of the HMG and SOX family members during evolution. Journal of molecular evolution 48: 517-527.

[326] Huang CR, Burns KH and Boeke JD (2012) Active transposition in genomes. Annual review of genetics 46: 651-675.

[327] Sela N, Kim E and Ast G (2010) The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. Genome biology 11: R59.

[328] Castaneda J, Genzor P and Bortvin A (2011) piRNAs, transposon silencing, and germline genome integrity. Mutation research 714: 95-104.

[329] Guenther UP, Yandek LE, Niland CN, Campbell FE, Anderson D, et al. (2013) Hidden specificity in an apparently nonspecific RNA-binding protein. Nature 502: 385-388.

[330] Jarrous N and Reiner R (2007) Human RNase P: a tRNA-processing enzyme and transcription factor. Nucleic acids research 35: 3519-3524.

[331] Kazantsev AV, Krivenko AA, Harrington DJ, Holbrook SR, Adams PD, et al. (2005) Crystal structure of a bacterial ribonuclease P RNA. Proceedings of the National Academy of Sciences of the United States of America 102: 13392-13397.

[332] Torres-Larios A, Swinger KK, Krasilnikov AS, Pan T and Mondragon A (2005) Crystal structure of the RNA component of bacterial ribonuclease P. Nature 437: 584-587.

[333] Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, et al. (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481: 389-393.

[334] R Development Core Team (2010) R: A language and environment for statistical computing Vienna, Austria: R Foundation for Staistical Computing.

[335] Bailey TL and Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology 2: 28-36.

[336] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic acids research 37: W202-208.

[337] Phan AT and Mergny JL (2002) Human telomeric DNA: G-quadruplex, i-motif and Watson-Crick double helix. Nucleic acids research 30: 4618-4625.

[338] Shrestha P, Xiao S, Dhakal S, Tan Z and Mao H (2014) Nascent RNA transcripts facilitate the formation of G-quadruplexes. Nucleic acids research 42: 7236-7246.

[339] Bugaut A and Balasubramanian S (2012) 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic acids research 40: 4727-4741.

[340] Lexa M, Kejnovsky E, Steflova P, Konvalinova H, Vorlickova M, et al. (2014) Quadruplex-forming sequences occupy discrete regions inside plant LTR retrotransposons. Nucleic acids research 42: 968-978.

[341] Kejnovsky E and Lexa M (2014) Quadruplex-forming DNA sequences spread by retrotransposons may serve as genome regulators. Mobile genetic elements 4: e28084.

[342] Millevoi S, Moine H and Vagner S (2012) G-quadruplexes in RNA biology. Wiley interdisciplinary reviews RNA 3: 495-507.

[343] Kikin O, D'Antonio L and Bagga PS (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic acids research 34: W676-682.

[344] Beard C, Hochedlinger K, Plath K, Wutz A and Jaenisch R (2006) Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. Genesis 44: 23-28.

BIOGRAPHICAL SKETCH

NAME	**POSITION TITILE**
Pavol Genzor	Ph.D Candidate
D.O.B.	**CITY, COUNTRY**
July 3, 1985	Dolny Kubin, Slovakia

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE	MM/YY	FIELD OF STUDY
Johns Hopkins University Baltimore, MD	Ph.D	05/2015	Molecular Biology
Saint Vincent College Latrobe, PA	B.S.	05/2008	Biology

HONORS

The Dupont Award for Excellence in Teaching Recipient (2010)
A. J. Palumbo Research Grant Recipient (2007)
International Student Grant Recipient (2004-2008)
Saint Vincent College Academic Scholarship Recipient (2004-2008)
Nominee for Who's Who Among Students in American Universities (2006)

TEACHING

Advanced Cellular Biology Teaching Assistant, Johns Hopkins University, Fall 2009
Advanced Developmental Biology Teaching Assistant, Johns Hopkins University, Fall 2010

PRESENTATIONS

"A unique HMG-box domain of mouse Maelstrom binds structured RNA but not double stranded DNA." Poster. RNA Biology Symposium. NIH. Bethesda. 2015

"Biochemical studies of a HMG-box domain of Maelstrom proteins." Poster. Germ Cell Meeting. Cold Spring Harbor. 2012

PUBLICATIONS

P. Genzor and A. Bortvin, (2015). *"A unique HMG-box domain of mouse Maelstrom binds structured RNA but not double stranded DNA."* PlosONE accepted

J. Castaneda, **P. Genzor**, G. W. van der Heijden, A. Sarkeshik, J. R. Yates, 3rd, N. T. Ingolia and A. Bortvin (2014). *"Reduced pachytene piRNAs and translation underlie spermiogenic arrest in Maelstrom mutant mice."* EMBO J 33(18): 1999-2019.

J. Castaneda*, **P. Genzor*** and A. Bortvin (2011). *"piRNAs, transposon silencing, and germline genome integrity."* Mutat Res 714(1-2): 95-104. * co-authors

A. Patel, J. N. McKnight, **P. Genzor** and G. D. Bowman (2011). *"Identification of residues in chromodomain helicase DNA-binding protein 1 (Chd1) required for coupling ATP hydrolysis to nucleosome sliding."* J Biol Chem 286(51): 43984-43993.